

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>5</sup> :  
C12N 15/63, 15/09, C12P 21/00

A1

(11) International Publication Number: WO 94/25609

(43) International Publication Date: 10 November 1994 (10.11.94)

(21) International Application Number: PCT/US94/04651

(22) International Filing Date: 28 April 1994 (28.04.94)

(30) Priority Data:  
08/054,730 28 April 1993 (28.04.93) US

(71) Applicant: HYBRITECH INCORPORATED [US/US]; 11095  
Torreyana Drive, San Diego, CA 92121 (US).

(72) Inventors: ANTELMAN, Douglas, Evan; 3808 Fallon Circle,  
San Diego, CA 92130 (US). WILSON, Barry, S.; 8442  
Highwood Drive, San Diego, CA 92119 (US).

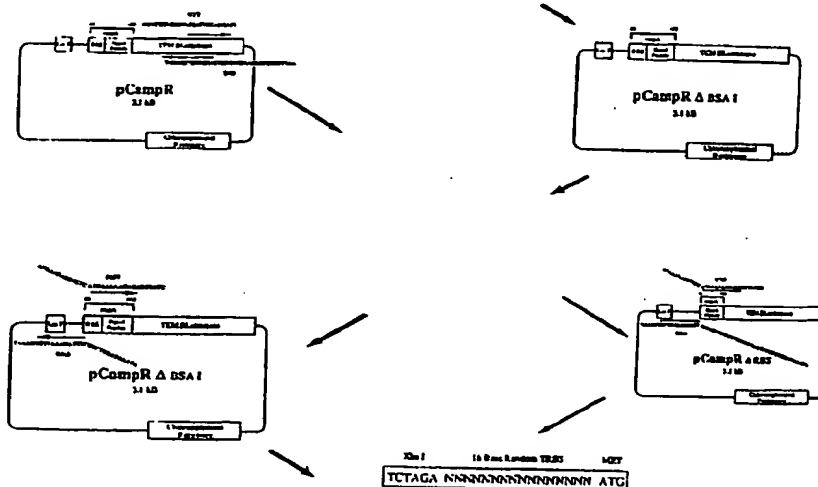
(74) Agents: LAMMERT, Steven, R. et al.; Barnes & Thornburg,  
1313 Merchants Bank Building, 11 South Meridian Street,  
Indianapolis, IN 46204 (US).

(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CN,  
CZ, DE, DK, ES, FI, GB, HU, JP, KG, KP, KR, KZ, LK,  
LU, LV, MD, MG, MN, MW, NL, NO, NZ, PL, PT, RO,  
RU, SD, SE, SI, SK, TJ, TT, UA, UZ, VN, European patent  
(AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC,  
NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA,  
GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.

(54) Title: METHOD FOR CREATING OPTIMIZED REGULATORY REGIONS AFFECTING PROTEIN EXPRESSION AND PROTEIN TRAFFICKING



(57) Abstract

This invention discloses a method for optimizing the production of a polypeptide in a cell comprising the steps of identifying a regulatory region within a nucleic acid sequence to be mutagenized; preparing a nucleic acid vector comprising that region and encoding at least one polypeptide regulated by said regulatory region; deleting that region from the vector; producing a pool of random oligonucleotides; using PCR to introduce one random oligonucleotide into the position previously occupied by the regulatory region into each of a plurality of vectors to generate a pool of mutagenized vectors; introducing the mutagenized vectors into a cell sample; assaying for the expression of the polypeptide in that sample; and selecting and isolating those cells exhibiting optimized polypeptide expression. The regulatory regions contemplated within the scope of this invention include signal and protein trafficking sequences, ribosome binding sites, promoters, and translational regulatory sequences.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LV	Latvia	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	MC	Monaco	TG	Togo
CZ	Czech Republic	MD	Republic of Moldova	TJ	Tajikistan
DE	Germany	MG	Madagascar	TT	Trinidad and Tobago
DK	Denmark	ML	Mali	UA	Ukraine
ES	Spain	MN	Mongolia	US	United States of America
FI	Finland			UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

-1-

METHOD FOR CREATING OPTIMIZED REGULATORY REGIONS  
AFFECTING PROTEIN EXPRESSION AND PROTEIN TRAFFICKING

5 This invention relates generally to methods for  
the mutagenesis of nucleic acid sequences and more  
specifically to mutagenesis strategies involving polymerase  
chain reaction-related technologies and to optimized  
regulatory sequences generated by these methods.

Gene expression can be regulated at various steps  
10 on the path from DNA to RNA to mature protein. The overall  
path is the same for prokaryotes and eukaryotes with  
regulation possible at the level of transcription of mRNA  
from the DNA template, translation of mRNA at the ribosome,  
and targeting of protein to appropriate locations within  
15 the cell or outside. However, the differences between  
prokaryotes and eukaryotes lead to a number of differences  
in how gene expression can be regulated in each.  
Intracellular membranes and organelles are critical  
components of eukaryotic cells, but are not found in  
20 prokaryotes. On the other hand, the potential for coupled  
transcription and translation in prokaryotes is not  
possible in compartmentalized eukaryotic cells. Eukaryotic  
genes have introns interrupting the coding sequences,  
prokaryotic genes do not. On the other hand, prokaryotes  
25 put coding sequences for multiple genes under a single  
promoter and transcribe them as a single polycistronic  
mRNA, but eukaryotes do not. Most eukaryotic genes require  
RNA splicing, capping, poly-adenylation, and transport from  
the nucleus to the cytoplasm before the mRNA becomes  
30 functional, but prokaryotes do not. In addition,  
eukaryotes require a more complicated process for protein  
targeting to distinguish proteins destined for different  
organelles, the cell membrane, and for secretion, as well  
as post translational modification of many eukaryotic  
35 proteins before they can be assembled or become functional.

SUBSTITUTE SHEET (RULE 26)

-2-

In **EUKARYOTES**, the expression of a particular gene can be regulated at transcription, RNA processing, RNA transport, translation, protein targeting, and protein activation/modification. Specific gene sequences are responsible for effects at each level and these can be modified to increase expression according to this invention, if eukaryotic expression is desired.

At the level of transcription, these sequences act directly, by interacting with DNA binding proteins. Examples include: **Promoters** which interact with transcription factors and RNA polymerase to initiate transcription. **Upstream promoter elements** which are generally near to the promoter and interact with DNA binding proteins to either improve or inhibit the ability of polymerase to interact with the promoter. The transcription of the genes encoding the DNA binding proteins described above and below is another level of regulation of specific eukaryotic genes. **Enhancers**, which are quite distant from the promoter, bind to other DNA binding proteins to enhance the level of transcription of the gene, although specific sequences are also present in enhancers which can inhibit transcription under certain conditions or in specific cell types. One skilled in the art can readily determine whether transcription is increased or decreased by specific sequences within upstream promoter elements or enhancers in a specific cell type. Thus, enhancers, promoters and upstream promoter elements can be modified to increase expression according to this invention.

At the level of **RNA processing**, primary transcripts are capped at their 5' ends, poly-adenylated at their 3' ends and have their introns spliced out (sometimes leading to different protein products due to **alternative RNA splicing**) prior to being transported through pores in the nuclear membrane to cytoplasm, where they associate

-3-

with ribosomes. Important RNA sequences which can affect the efficiency of this process include the poly-A addition site, the Cap site, and the splice acceptor and donor sites, and DNA sequences encoding these sites can be  
5 modified to increase expression according to this invention.

To participate in translation, a eukaryotic mRNA must be transported out of the nucleus. Specific recognition signals on the mRNA are believed necessary for  
10 transit of mRNA through the nuclear pore. Other signals may retain certain mRNAs within the nucleus, where they are degraded. As many as half of the unspliced heterologous nuclear RNAs produced from primary transcripts are degraded without ever leaving the nucleus, offering a potential for  
15 regulation at another level, the level of RNA transport. DNA sequences encoding these RNA signals could be modified to increase expression according to this invention as well.

Not all mRNAs that escape the nucleus are translated into protein. Eukaryotic mRNAs vary in  
20 susceptibility to degradation. Binding to ribosomes decreases their degradation. Important sequences that can be modified to control expression at the level of translation are those recognized by specific translation repressor proteins, which bind to the 5' end of some mRNAs  
25 and block translation initiation. For those mRNAs which are not blocked, current data suggests that initiation of translation is by the so-called "scanning model" of initiation. A ribosome component (40S) binds initially to the 5' end of an mRNA then migrates 3' until it finds an  
30 initiation site where the other portion of the ribosome joins it to form the 80S initiation complex. The Kozak sequences surrounding the translational start site determine whether a particular mRNA will be efficiently translated. The presence of an Adenine (A) at the -3  
35 position has as much as a twenty-fold impact on

SUBSTITUTE SHEET (RULE 26)

-4-

translation. In general, the translation rate is most affected by the nucleotide three prior to the initiator codon and the one immediately after it [best if a Guanine (G)]. The DNA sequence encoding binding sites for translation repressor proteins and for the region around the initiator can be modified to increase expression according to this invention.

At the level of protein targeting, eukaryotic proteins have a diversity of target locations within the cell including the nucleus, nuclear membrane, cytoplasm, mitochondria, mitochondrial inner membrane, mitochondrial outer membrane, endoplasmic reticulum, golgi apparatus, lysosomes, lysosomal membranes, endoplasmic reticulum membrane, and cell membrane. In addition, secreted proteins are targeted to the exterior of the cell. The process begins, as for prokaryotes, with a signal sequence directing the nascent peptide and its ribosomes to a membrane. But in this case, the membrane is the rough endoplasmic reticulum (RER). If the protein is to be secreted, it is translated into the RER and begins a journey through different membrane bound organelles and vesicles to the cell membrane where it is extruded to the outside by exocytosis. Proteins destined to be membrane-bound are also translated into the RER, but do not completely transit the rough endoplasmic reticulum membrane (RERM), becoming integrated into it. This membrane fuses with that of the Golgi apparatus or of vesicles which, in turn, bud off and fuse with the destination membrane where the protein either becomes bound or transits further to its final target. Targeting to the proper site is almost certainly facilitated by identifiable target signals within the sequence of the protein that direct the protein to the proper destination. Thus, the sequences that can be modified to increase expression according to this invention include those encoding the signal sequence, transmembrane

-5-

sequences, cytoplasmic anchor sequences, and the putative specific membrane targeting sequences.

In addition, numerous secreted proteins, including immunoglobulins, consist of multiple chains that must be assembled prior to secretion. In immunoglobulins, for example, the heavy chains cannot be secreted if they are not bound to light chains. The heavy chains that do not bind to light chains, either because of improper folding or an absence of free light chains, remain bound to a protein in the RER called BiP, which prevents their transit to the Golgi apparatus. Thus, it is clear that additional regulatory elements that can be modified according to the method of this invention to increase expression are those which encode structural features that facilitate or impede correct folding of a peptide or its association with other chains necessary to form the mature protein product.

At the level of protein activation/modification, many eukaryotic proteins are synthesized as inactive or partially active precursors that become activated by proteolytic cleavage or other modifications (e.g., phosphorylation). Although, not required for expression in many cases, secretion cannot occur without proteolytic cleavage of the signal peptide. Inefficiency in this process can delay or block secretion of functional protein. DNA sequences encoding proteolytic cleavage sites, phosphorylation sites, glycosylation sites, or other sites for post-translational modification can be modified according to this invention to yield increased expression of functional protein.

In PROKARYOTES, there are fewer stages where regulation of gene expression can occur, and prokaryote specific elements and associated proteins are involved. Regulation does take place at the levels of transcription, translation, and protein targeting, as for eukaryotic

SUBSTITUTE SHEET (RULE 26)

-6-

genes, but RNA processing and RNA transport are totally absent in prokaryotes. There are also far fewer known cases of post-translational modification of proteins in bacteria than in eukaryotes. Of special importance, there are very few authenticated cases of glycosylated bacterial proteins, although proteins are modified in other ways (e.g., fatty acylation). No clear picture of regulation at the level of post-translational modification has emerged. Like eukaryotes, prokaryotes offer a number of specific DNA sequences which can be modified according to this invention to increase expression.

At the level of transcription in prokaryotes identifiable DNA sequences interact with DNA binding proteins. The promoter binds RNA polymerase and, with the release of the  $\sigma$  subunit, initiation of transcription occurs. Unlike in eukaryotes, distant enhancer sequences have not been seen in prokaryotes. Different promoter sequences have different affinities for RNA polymerase. Transcription initiation can be positively regulated by binding of a catabolite gene activator protein (CAP) to a site immediately upstream of the promoter and, in many genes, can be negatively regulated by binding of a repressor protein to a site immediately downstream of the promoter called the operator. The level of transcription of these DNA binding proteins can affect the transcriptional level of the regulated gene. As one skilled in the art will appreciate, binding sites can be modified both in the target gene and in the gene encoding the repressor or CAP using the present invention to increase or decrease the level of transcription of prokaryotic genes.

Sites amenable to modification to increase expression at the level of translation also exist in prokaryotic DNA, but are distinct from those in eukaryotes. A region originally known as the Shine and Dalgarno

-7-

sequence or more recently as the ribosome binding site found on the 5' end of prokaryotic mRNAs is complementary to the 16S ribosomal RNA. Similar to eukaryotic initiators, an Adenine (A) is preferred at the -3 position.

5 The Shine and Dalgarno interaction is unique to prokaryotes and is unlike the scanning model used in eukaryotes to initiate translation described above. The DNA sequence which encodes the ribosome binding site in mRNA can be modified according to this invention to increase

10 expression.

There are proteins which bind to sites on specific mRNAs to inhibit their translation (e.g., ribosomal proteins). In addition, transcription can be affected by 5' coding sequences which can lead to premature

15 termination of transcription based on charged tRNA availability. This system, called attenuation, is seen in genes for amino acid biosynthetic pathways, but is unique to prokaryotes because it requires coupled transcription and translation, which can occur only in cells having no

20 nuclear membrane separating the two processes. Placement of the gene within the operon also has an effect on translation of specific genes. The 5' genes are more efficient in initiation than those at the 3' end of the polycistronic message. The "polarity effect" is amplified

25 by the fact that binding to ribosomes actually delays mRNA degradation. The DNA sequences which encode protein sequences responsible for attenuation can be modified according to this invention to increase expression.

At the level of protein targeting, as with

30 eukaryotes, specific targeting sequences are required for different types of targeting. These sequences encode portions of the protein with hydrophobic or hydrophilic properties corresponding to the protein's point of association with the membrane. The most common sequence is

35 the N-terminal signal sequence, which is responsible for

**SUBSTITUTE SHEET (RULE 26)**

-8-

targeting the protein and its associated ribosomes to the cell membrane. Internal sequences determine if portions of the protein will reside within or outside of the lipid bilayer of the membrane or if the protein will be secreted  
5 into the periplasmic space (in gram negative) or directly into the media (in gram positive bacteria).

The efficiency of secretion depends upon the signal sequence as well as the cellular machinery required for transport. This includes peptidases for removing the  
10 signal peptide and bacterial cell membrane proteins, both of which have been implicated as potential signal peptide receptors. A Signal Recognition Particle (SRP) analogue has recently been confirmed for bacteria (Luirink J., et al. Nature, 359:741-743, 1992. Another significant  
15 difference between prokaryotes and eukaryotes is the limited number of target membranes in bacteria.

Due to all of these differences in the mechanisms of expression between eukaryotes and prokaryotes heterologous expression of eukaryotic genes in prokaryotic  
20 cells faces a number of hurdles. First, all introns must be removed or truncated products will be produced. Second, eukaryotic upstream promoter elements and enhancers are not only non-functional in prokaryotes but can be deleterious. Third, eukaryotic signals, while occasionally functional,  
25 are never optimized for prokaryotes, and often are totally ineffective. Often fortuitous prokaryote-like signal sequences are used that are distinct from the native eukaryotic signal. Fourth, the absence of functional prokaryotic ribosome binding sequences often leads to  
30 little or no translation of the eukaryotic message. Fifth, the absence of prokaryotic codon usage can slow translation. Finally, the DNA sequence encoding an mRNA for a eukaryotic signal could be completely misinterpreted in a prokaryotic expression system. For example, it could  
35 be recognized by a prokaryotic DNA binding protein, or the

SUBSTITUTE SHEET (RULE 26)

-9-

mRNA could experience inhibitory folding or be more rapidly degraded than in the native cell. Because such miscommunication depends in part upon the sequence of the gene to be expressed, it is possible that different genes inserted into the same vector in the same heterologous expression cell will attain different levels of expression. For example, a leader sequence optimized to attain the highest level of expression of one eukaryotic protein gene in *E. coli* may not be the leader sequence that will optimize expression of a different eukaryotic protein gene in *E. coli*.

The present invention overcomes these difficulties by providing a means to customize heterologous expression for any particular desired protein. The present method assays the combined effect of mutations made in the DNA of one or more particular control sequences upon all steps taken during the whole process of heterologous expression leading up to secretion of a particular desired protein. Therefore, the present method provides a means to customize modifications to discrete sequences in eukaryotic DNA which impact expression of the desired protein in the heterologous host cell.

For a general review of the differences in eukaryotic and prokaryotic gene expression and protein transport see: Watson, JD, et al. Molecular Biology of the Gene 4th edition. Menlo Park, Benjamin Cummings Press, pp. 359-618 (1987); Alberts, et al. Molecular Biology of the Cell 2nd ed. New York, Garland Publ., pp. 551-612 (1989); Reznikoff, W. and Gold, L. eds. Maximizing Gene Expression Boston, Butterworths Publ., pp. 1-34, 195-286 (1986); Pugsley, A.P. Protein Targeting San Diego, Academic Press, pp. 1-168 (1989) and Watson, et al. Recombinant DNA: A Short Course New York, Freeman Publ. pp. 45-57. For a review of protein translocation across the *E. coli* plasma

SUBSTITUTE SHEET (RULE 26)

-10-

membrane see Wichner, et al. *Annu. Rev. Biochem.* 60:101-24, 1991 and Bieker, et al. *Trends in Gen.* 6:329-333, 1990.

The references listed in this application are incorporated herein by reference to the extent that they supplement, explain, or provide a background for each methodology, technique and/or composition employed herein.

This invention relates to methods for optimizing the expression of polypeptide in a cell and to methods for creating and isolating novel regulatory sequences for the expression of polypeptide in a cell. The regulatory sequences contemplated within the scope of this invention include, but are not limited to, signal sequences, ribosome binding sites, promoter sequences, translational regulatory sequences, transcription regulatory sequences, protein trafficking sequences, enhancer sequences, and the like.

In a preferred aspect of the invention, a method is provided for optimizing the production of polypeptide in a cell comprising the steps of (a) identifying at least one regulatory region within a nucleic acid sequence to be mutagenized; (b) preparing a nucleic acid vector comprising the regulatory region and a nucleic acid sequence encoding at least one polypeptide regulated by the regulatory region; (c) deleting the regulatory region from the vector; (d) producing a pool of random oligonucleotides; (e) using a polymerase chain reaction to introduce at least one random oligonucleotide into the position previously occupied by the regulatory region in a plurality of vectors to generate a pool of mutagenized vectors; (f) introducing the mutagenized vectors into a cell sample; (g) assaying for the expression of the polypeptide in the cell sample; (h) selecting cells exhibiting optimized polypeptide expression; and (i) isolating optimized polypeptide from the cells of step (h). Preferably the regulatory region is located within the nucleic acid sequence encoding polypeptide and in one preferred embodiment the regulatory

SUBSTITUTE SHEET (RULE 26)

-11-

region is in a translated portion of the nucleic acid sequence encoding polypeptide. In another preferred embodiment, the regulatory region is outside of the nucleic acid sequence encoding polypeptide. The method is useful  
5 for optimizing regulatory regions selected from the group consisting of signal sequences, ribosome binding sites, promoter sequences, translational regulatory sequences, transcription regulatory sequences and protein trafficking sequences. In a preferred embodiment of this invention,  
10 the nucleic acid sequence encodes an antibiotic resistant gene and the selection step consists essentially of growing the cell sample in the presence of an antibiotic. In another aspect of this embodiment, the nucleic acid sequence encoding polypeptide encodes a selectable marker  
15 and in yet another aspect of this embodiment, the nucleic acid sequence encoding polypeptide encodes a fusion protein. In a preferred embodiment the nucleic acid sequence encoding polypeptide is derived from a eukaryotic cell and the cell sample is prokaryotic.

20 In a preferred method for optimizing the production of polypeptide in a cell, the method additionally comprises the steps of introducing the nucleic acid vector into a second cell sample, expressing the polypeptide encoded by the vector in the second cell  
25 sample, and measuring the level of polypeptide expression in the second cell sample. This method optionally includes the step of selecting cells from the first cell sample that exhibit optimized polypeptide expression relative to the measuring step. Once optimized cells are selected, it is  
30 contemplated that the nucleic acid sequence of the random oligonucleotide corresponding to the mutagenized regulatory region producing optimized polypeptide expression can be identified.

In another preferred embodiment of this  
35 invention, a method is provided for creating and isolating

-12-

novel signal sequences comprising the steps of: (a) identifying a signal sequence within a nucleic acid sequence encoding a polypeptide; (b) preparing a nucleic acid vector comprising the nucleic acid sequence encoding the polypeptide; (c) introducing the vector into a first cell sample and expressing the polypeptide in the first cell sample; (d) measuring the level of polypeptide expression in the first cell sample; (e) deleting the signal sequence from the vector; (f) producing a pool of random oligonucleotides; (g) using a polymerase chain reaction to introduce at least one of the random oligonucleotides into the position previously occupied by the signal sequence in a plurality of vectors to generate a pool of mutagenized vectors; (h) introducing the mutagenized vectors into a second cell sample; (i) assaying for the expression of the polypeptide in the second cell sample; (j) selecting cells exhibiting optimized polypeptide expression relative to step (d); and (k) determining the nucleic acid sequence of the regulatory region contained in the mutagenized vector located in the cells of step (j). In one embodiment the optimized level of polypeptide expression is a level of expression greater than or equal to the level of expression of the polypeptide in step (d).

Preferably the nucleic acid vector encodes an antibiotic resistant gene and the selection step additionally consists essentially of growing the cell sample in the presence of an antibiotic. The nucleic acid sequence encoding polypeptide preferably encodes a selectable marker and the polymerase chain reaction is preferably an enzymatic inverse polymerase chain reaction. In one embodiment the nucleic acid sequence encoding polypeptide encodes a fusion polypeptide and in a particularly preferred embodiment a portion of the fusion protein is derived from  $\beta$ -lactamase. In another embodiment

SUBSTITUTE SHEET (RULE 26)

-13-

a portion of the fusion protein is derived from an antibody. Preferably the nucleic acid sequence encoding polypeptide is derived from a eukaryotic cell and the first and second cell samples are prokaryotic. Still more  
5 preferably, the cell samples are *E. coli*.

In yet another preferred embodiment of the method for creating and isolating novel signal sequences, the random oligonucleotides are biased and preferably contain at least one positively charged amino acid at the N-  
10 terminus, a stretch of at least 8 hydrophobic amino acids and a small amino acid such as alanine, glycine, or valine positioned at the C-terminus.

In another preferred method of this invention, a method is disclosed for optimizing polypeptide expression  
15 in a cell by performing random mutagenesis on a regulatory region regulating polypeptide expression from a nucleic acid vector wherein the improvement comprises deleting a region of nucleic acid to be mutagenized, isolating the vector containing the deletion and replacing the region to  
20 be mutagenized with a random nucleic acid sequence.

In a preferred method of this invention a method is provided for creating and isolating novel ribosome binding sites, comprising the steps of: (a) identifying a ribosome binding site within a nucleic acid sequence; (b)  
25 preparing a nucleic acid vector comprising the nucleic acid sequence containing a ribosome binding site operably linked to a nucleic acid sequence encoding a polypeptide; (c) introducing the vector into a first cell sample and expressing the polypeptide in the first cell sample; (d)  
30 measuring the level of polypeptide expression in the first cell sample; (e) deleting the nucleic acid sequence containing the ribosome binding site from the vector; (f) producing a pool of random oligonucleotides; (g) using a polymerase chain reaction to introduce at least one of the  
35 random oligonucleotides into the position previously

SUBSTITUTE SHEET (RULE 26)

-14-

occupied by the nucleic acid sequence containing the ribosome binding site in a plurality of vectors to generate a pool of mutagenized vectors; (h) introducing the mutagenized vectors into a second cell sample; (i) assaying  
5 for the expression of the polypeptide in the second cell sample; (j) selecting cells exhibiting optimized polypeptide expression relative to step (d); and determining the nucleic acid sequence containing the ribosome binding site located in the mutagenized vector  
10 introduced into the cells of step (j). In a preferred embodiment of this method the nucleic acid sequence encodes an antibiotic resistance gene and the selection step consists essentially of growing the cell samples in the presence of an antibiotic. Preferably, the nucleic acid  
15 sequence encoding polypeptide encodes a selectable marker. In one embodiment the nucleic acid sequence encoding polypeptide is a fusion protein and preferably the fusion protein is derived from an antibody and in another preferred embodiment the fusion protein is derived from  
20  $\beta$ -lactamase.

In another preferred aspect of this invention, polypeptide signal sequences are disclosed. These signal sequences correspond to SEQ ID NO: 13, SEQ ID NO: 16 and SEQ ID NO: 14 and to polypeptide signal sequences  
25 containing at least contiguous amino acid 10-mers thereof. In another preferred aspect of this invention, nucleic acid sequences are disclosed corresponding to ribosome binding sites identified as SEQ ID NO: 35, SEQ ID NO: 37, and SEQ ID NO: 38 and to ribosome binding sites corresponding to at  
30 least contiguous 5-mers thereof.

In yet another preferred embodiment of this invention, a method is provided for creating and isolating novel signal sequences, comprising the steps of: (a) identifying a signal sequence within a nucleic acid  
35 sequence encoding a polypeptide; (b) preparing a nucleic

-15-

acid vector comprising the nucleic acid sequence encoding a polypeptides, wherein the polypeptide is a fusion protein having a C-terminus corresponding to  $\beta$ -lactamase; (c) deleting the signal sequence from the vector; (d) producing  
5 a pool of random oligonucleotide suitable for an enzymatic inverse polymerase chain reaction; (e) using an enzymatic inverse polymerase chain reaction to introduce the random oligonucleotides into the position in the vector previously occupied by the signal sequence to generate a pool of  
10 mutagenized vectors; (f) introducing the mutagenized vectors into a first cell sample of *E. coli*; (g) assaying for the expression of the fusion protein in a first cell sample in the presence of ampicillin; and (h) selecting cells exhibiting optimized polypeptide expression. In one  
15 embodiment of this method, the method additionally comprises the steps of introducing the vector comprising the nucleic acid sequence encoding a polypeptide into a second cell sample of *E. coli*, expressing the polypeptide in the second cell sample and measuring the level of  
20 polypeptide expression in the second cell sample. In one embodiment this method includes the added step of selecting cells from the first cell sample that exhibit optimized polypeptide expression relative to the level of polypeptide expression observed in the measuring step. Preferably, the  
25 optimized polypeptide expression is a level of polypeptide expression greater than or equal to the level of polypeptide expression obtained from the measuring step. In another preferred embodiment, the method additionally comprises the step of determining the nucleic acid sequence  
30 of the random oligonucleotide introduced into the position in the vector, previously occupied by the signal sequence, isolated from cells identified by the selecting step. In yet another preferred embodiment of the invention the concentration of ampicillin is at least 30  $\mu\text{g/ml}$  and the

-16-

nucleic acid sequence encoding polypeptide is a sequence encoding a single-chain antibody.

In another method of this invention, a method is disclosed for identifying a novel nucleic acid sequence encoding a protein trafficking signal that directs a polypeptide to a desired location in a cell, comprising the steps of: (a) identifying a region of nucleic acid containing at least one protein trafficking signal to be mutagenized; (b) preparing a nucleic acid vector comprising the protein trafficking signal sequence to be mutagenized and a nucleic acid sequence encoding at least one polypeptide; (c) deleting the protein trafficking signal sequence from the vector; (d) producing a pool of random oligonucleotides; (e) using a polymerase chain reaction to introduce at least one of the random oligonucleotides into the position previously occupied by the protein trafficking signal in a plurality of vectors to generate a pool of mutagenized vectors; (f) introducing the mutagenized vectors into a cell sample; (g) assaying for the location of the polypeptide in the cell sample; (h) selecting cells expressing the polypeptide in the desired cell location; and (i) determining the nucleic acid sequence of the protein trafficking signal contained in the mutagenized vector located in the cells of step (h).

In a preferred embodiment of the trafficking method of this invention, the protein trafficking sequence to be mutagenized is located in the nucleic acid sequence encoding the polypeptide and in another embodiment the polymerase chain reaction is the enzymatic inverse polymerase chain reaction and the nucleic acid sequence encodes an antibiotic resistant gene and the selection step consists essentially of growing the cell sample in the presence of an antibiotic. Preferably the desired cell location is extracellular.

-17-

In another aspect of this invention, a method is provided for creating and isolating novel regulatory sequences useful for optimizing the expression of a recombinant polypeptide in prokaryotic cells comprising the

5 steps of: (a) identifying at least one regulatory region within a nucleic acid sequence to be mutagenized; (b) preparing a nucleic acid vector suitable for expressing polypeptide in a prokaryotic cell, the nucleic acid vector comprising the regulatory region and a nucleic acid

10 sequence encoding at least one polypeptide operably linked to the regulatory region; (c) introducing the vector into a first prokaryotic cell sample and expressing the polypeptide encoded by the vector in the first cell sample; (d) measuring the level of polypeptide expression in the

15 first prokaryotic cell sample; (e) deleting the regulatory region from the vector; (f) producing a pool of random oligonucleotides; (g) using a polymerase chain reaction to introduce at least one random oligonucleotide into the position previously occupied by the regulatory region in a

20 plurality of vectors to generate a pool of mutagenized vectors; (h) introducing the mutagenized vectors into a first prokaryotic cell sample; (i) assaying for the expression of the polypeptide in the second prokaryotic cell sample; (j) selecting cells exhibiting optimized

25 polypeptide expression relative to step (d); and (k) determining the nucleic acid sequence of the mutagenized regulatory region located within the vector introduced into the cells of step (j).

In a preferred embodiment of this method, the

30 regulatory region is selected from the group consisting of a signal sequence, a ribosome binding site, a promoter sequence, a translational regulatory sequence, a transcription regulatory sequence and a protein trafficking sequence. In one embodiment the nucleic acid sequence

35 encoding polypeptide encodes an antibiotic resistant gene

**SUBSTITUTE SHEET (RULE 26)**

-18-

and the selection step additionally consists of growing the second prokaryotic cell sample in the presence of an antibiotic. Preferably the first and second cell samples are derived from the same cell type.

5                Figure 1 diagrams the assembly of the single chain antibody  $\beta$ -lactamase fusion protein and the incorporation of the fusion protein with the Omp A signal sequence into the expression vector pCCHAScl-ampRV2.

                 Figure 2 illustrates the strategies followed to  
10 delete the Omp A regulatory region, to incorporate the random oligonucleotides, and to generate a library of random signal peptide mutants.

                 Figure 3 is a diagram of plasmid pCLA3ampR that was used to reclone identified signal sequences to test for  
15 signal sequence effectiveness.

                 Figure 4 is a comparison of the hydrophobicity and the alpha and beta region analysis of two novel signal peptide sequences identified using the methods of this invention as compared with known signal peptide sequences.

20                Figure 5 illustrates the construction of variable length signal peptide libraries using the methods of this invention.

                 Figure 6 is a diagram of the plasmid modifications used to produce the plasmid containing the  
25 ribosome binding site library.

                 Figure 7 is a diagram of plasmid pGCEMK.

                 Figure 8 is a diagram of plasmid pNCEMG1.

                 Figure 9 illustrates sequences of some of the oligonucleotides containing the ribosome binding site and  
30 signal peptide sequences identified using the methods of this invention.

                 This invention provides methods useful for creating novel regulatory sequences that affect recombinant protein expression within either a prokaryotic or  
35 eukaryotic cell. In particular, these methods

SUBSTITUTE SHEET (RULE 26)

-19-

advantageously facilitate the optimization of recombinant protein expression in a cell. Using the techniques of this invention, modifications can be incorporated into more than one regulatory region and the cumulative effect of these mutations can be assessed by isolating clones expressing the desired level of recombinant protein expression.

The term "polypeptide" and "protein" are used interchangeably throughout this text.

The term "heterologous protein expression" is used herein to refer to protein that is not native to the host cell expressing the heterologous protein.

The term "optimized protein expression" is used herein to refer to the cumulative conditions that provide an optimal level of protein expression for a particular protein in a particular cell system. Under one set of laboratory conditions, optimized protein expression may refer to the highest available level of protein expression, while in another set of laboratory conditions, optimized protein expression may refer to low level protein expression because for that particular application low level protein expression is preferred. In yet another set of laboratory conditions optimized protein expression may refer to the level of protein expression that can coexist with cell life in situations where, under standard conditions, the protein would be cytotoxic to the cells.

The term "operably linked" is used to mean regulatory regions that ultimately influence, or effect the level of recombinant protein expression from a nucleic acid vector.

The term "regulatory region is outside of a nucleic acid sequence encoding polypeptide" is used herein to indicate that the regulatory region of interest is not located within the coding region. For this purpose, the term coding region is used to denote that region of DNA

SUBSTITUTE SHEET (RULE 26)

-20-

that corresponds to RNA beginning with a methionine codon and ending with a polyadenylation signal.

While the methods of this invention are suitable for the optimization of recombinant protein sequences in either prokaryotes or eukaryotes, it is contemplated that these methods are particularly useful for the optimization of heterologous protein sequences in either prokaryotes or eukaryotes. In particular, these methods are useful for optimizing the expression of eukaryotic protein in prokaryotic cells.

Expression of eukaryotic protein in a prokaryotic system is often commercially desirable. Prokaryotic cells typically have shorter doubling times than eukaryotic cells and they are easier and less expensive to grow in large quantity than eukaryotes. However, optimizing expression of a eukaryotic protein in a prokaryotic cell such as *E. coli* has heretofore been an inefficient process. Typically a eukaryotic gene sequence encoding protein that is incorporated into a prokaryotic expression vector is expressed at relatively low levels or at levels that result in prokaryotic cell death. Given the multitude of differences in protein expression between prokaryotes and eukaryotes, as outlined in the background of this invention, it is often difficult to identify exactly why a particular eukaryotic sequence encoding protein is either not efficiently expressed or is bacteriocidal. The methods of this invention do not require a systematic assessment of each individual factor that can effect protein synthesis. Rather, the methods of this invention employ a novel random mutagenesis and selection strategy that advantageously permits one of skill in the art to identify clones that yield optimized levels of heterologous protein expression by assessing the cumulative effect of all of the changes incorporated into the expression vector. Thus, by the method of this invention, it is possible to select clones

SUBSTITUTE SHEET (RULE 26)

-21-

having mutagenized regulatory sequences individually tailored to optimize expression of any given protein in any given expression cell.

Most mutagenesis techniques involve the systematic replacement of one or more nucleotides with other known nucleotides. A variety of site-directed mutagenesis strategies are known in the art and are commercially available in kit format (see BioRad, Richmond, CA; Stratagene, La Jolla, CA; or Invitrogen, San Diego, CA.) Here, using site-directed mutagenesis, one begins with a known, efficient regulatory region and improves it by introducing single or multiple point mutations. In this procedure, the mutagenized regulatory region may or may not be biased according to what is known about the chemical properties of the wild-type regulatory region. For site-directed mutagenesis techniques used to optimize a prokaryotic signal peptide through the modification of a hydrophobic region of the signal peptide see Kendall, et al. *Nature* 321:706-708, 1986. For a review of site-directed mutagenesis strategies for prokaryotic signal peptides see Ryan, et al. *J. Biol. Chem.* 261:3389-3395, 1986; Park, et al. *Agric. Biol. Chem.* 55:1745-1750, 1991; vigal, et al. *Mol. Gen. Genet.* 231:88-96, 1991; Borchert, et al. *J. Bacteriol.* 173:276-282, 1991; Morioka-Fujimoto, et al. *J. Biol. Chem.* 266:1728-1732, 1991. Similarly, cassette mutagenesis strategies are also known in the art. These strategies involve removing a section of DNA between two restriction sites and replacing it with a different DNA sequence that is bounded by the same restriction sites (Richards, J.H., (1991) "Cassette Mutagenesis" in Directed Mutagenesis, A Practical Approach, M.J. McPherson ed. IRL Press). Chou, et al. (*J. Biol. Chem.* 265:2873-2880, 1990) studied the hydrophobicity requirements of the prokaryotic signal sequence using cassette mutagenesis. Ngsee, et al. used cassette mutagenesis in eukaryotic cells to study the

SUBSTITUTE SHEET (RULE 26)

-22-

effect of signal sequence mutants on protein expression (*Mol. Cell. Biol.* 9:3400-3410, 1989). These techniques replace identified sequences with known alternatives. However, unlike the mutagenesis strategies of this invention, these traditional site-directed and cassette mutagenesis techniques are not efficient methods for creating novel functional sequences.

In the practice of this invention, random or biased mutagenesis techniques are useful methods for generating a pool of novel mutants. Regions of oligonucleotides, or whole oligonucleotides, are synthesized using methods that permit the random incorporation of nucleotides. Random mutagenesis combined with good selection systems can be used to identify functional regulatory sequences lacking any predisposed bias to previously identified sequences. The collection of random sequences is incorporated into the gene sequence in place of the native sequence.

For purposes of this application, the term random mutagenesis will be used to encompass biased mutagenesis techniques. Biased mutagenesis is a type of random mutagenesis in which pools of oligonucleotides are prepared to accommodate the incorporation of more than one type of nucleotide in a given location within a nucleic acid sequence. Biased mutagenesis techniques are particularly useful for generating pools of mutants that maintain the consensus patterns of charge, hydrophobicity, hydrophilicity, etc. within a particular region of protein encoded by a nucleic acid sequence. In this way the chemical characteristics of a molecule can be maintained while the actual nucleic acid sequence and amino acid sequence of an encoding protein is varied. The application of biased mutagenesis to regulatory regions has the advantage that optimized sequences can be identified that are chemically biased toward native sequences but reflect a

-23-

wider variation than what has previously been identified in the art. Some of these sequences may direct protein expression more efficiently than other sequences heretofore identified.

5 Strategies of random mutagenesis to select clones containing functional mutations is complicated by the contaminating background level of clones containing nonmutated sequences. The presence of nonmutated sequences complicates selection and screening strategies by  
10 decreasing the variability within a library and by increasing the number of false positive clones. A large number of the clones produced by traditional random mutagenesis techniques are false positive because they contain the functional nonmutated sequence. Since the  
15 presence of novel optimized sequences from a random or biased pool of nucleotides is a relatively rare event, contaminating protein expression directed from nonmutated sequences can easily block the signal of the rarer, optimized sequence.

20 In this invention the drawback of non-mutated clones is avoided by substantially removing the wild-type sequence and replacing it with the randomly mutagenized sequence before selection of clones for function characteristics.

25 As noted, this invention is particularly useful for generating novel regulatory sequences associated with protein expression. The eukaryotic regulatory sequences contemplated within the scope of this invention include, but are not limited to promoters, upstream promoter  
30 elements, enhancers, transcriptional regulatory elements, translational regulatory proteins such as translation repressor proteins, signal sequences, protein targeting sequences, chaperon proteins (these binding proteins coat polypeptide chains as they emerge from ribosomes or  
35 membranes, preventing aggregation reactions and premature

SUBSTITUTE SHEET (RULE 26)

-24-

folding (Flynn, et al. *Nature* 353: 726, 1991) and sequences directing post-translational modifications and the like. The prokaryotic regulatory sequences contemplated within the scope of this invention include, but are not limited to, promoters, transcription regulatory sequences, translation regulatory sequences including ribosome binding sites, sequences associated with attenuation, protein targeting sequences, signal sequences, and the like.

Thus, the regulatory sequences may be in any number of locations relative to the gene sequence encoding protein, for instance, within the translated region of a coding sequence, within a transcribed region of a coding sequence, or external to the coding sequence. For example, the signal sequence for both eukaryotic and prokaryotic cells is positioned at the amino terminus of the nucleic acid sequence encoding protein. Since the primary sequence of the signal peptide has a direct effect on the translocation of the associated protein (Inouye, et al. *Proc. Natl. Acad. Sci., USA* 74: 1004-1008, 1977), the signal peptide is a candidate regulatory sequence suitable for the methods of this invention. Example 1 illustrates the application of the methods of this invention to the generation of novel signal sequences. As a second example to illustrate that the methods of this invention can be applied to a variety of regulatory sequences, Example 3 provides an exemplary strategy useful for generating novel ribosome binding sites.

The regulatory sequences created by the methods of this invention may be of any length. For example, while most signal sequences fall between a range of between 18-24 amino acids, it is contemplated that novel optimized signal sequences could be created in a variety of lengths. Example 2 provides methods for generating libraries of signal peptides having lengths of 12, 16 or 20 amino acids. Similarly these methods could be applied to other

-25-

regulatory regions such as ribosome binding sites, protein targeting sequences and the like.

#### Identification of a Regulatory Region to be Mutagenized

5           As a first step for practicing the methods of this invention, the position of the regulatory region to be mutagenized is identified. Regulatory regions contemplated within the scope of this invention have been discussed above. The relative position of the regulatory region can  
10 be identified within a particular nucleic acid sequence by studying the consensus data for that particular regulatory region available from the prior art. There exists in the literature a wealth of information concerning the consensus positions of known regulatory regions. For example,  
15 regions encoding signal peptides have been localized for both prokaryotic and eukaryotic genes as have regions encoded on mRNA which are required for ribosome binding (Protein Targeting. supra, pp. 52-61; Kozak, *Microbiol. Rev.* 47:1-45, 1983; and Kozak, *Nucl. Acids Res.* 15: 8125-  
20 8147, 1987). The precise identification of the position of a regulatory region within or in association with a nucleic acid sequence encoding protein is not necessary. Rather, any particular region of the nucleic acid containing a putative regulatory region postulated to exist within a  
25 nucleic acid sequence can be subjected to the mutagenesis techniques of this invention. These putative regulatory regions may be postulated in the art or they may be identified by trial and error using site-directed mutagenesis, or other mutagenesis techniques known in the  
30 art.

#### Preparing the Nucleic Acid Vector

Once the regulatory region to be mutagenized is identified, it is then incorporated into a nucleic acid  
35 vector together with a nucleic acid sequence encoding at

-26-

least one protein under the control of the regulatory region to be mutagenized. Any suitable expression vector may be used in this invention and exemplary vectors are provided in the Examples below. Those with skill in the art will appreciate that the choice of vector is limited to those vectors capable of directing expression of the nucleic acid sequence encoding protein and to those vectors that can incorporate and support the function of the regulatory region to be mutagenized. Further, the choice of vector is limited by the cell type selected; not all vectors and not all regulatory elements necessary for recombinant protein expression function in all cell types. As a general rule, eukaryotic expression vectors are suitable for protein expression in eukaryotes and prokaryotic expression vectors are suitable for prokaryotes. Both types of vectors are commercially available and those with skill in the art of molecular biology will be able to select the appropriate vectors suitable for recombinant protein expression within a given cell type. In addition, the following vectors used in the Examples of this invention (pCLA3ampR; pCCHAScl-ampRV2; and pCampR) have been deposited with the American Type Culture Collection (Bethesda, Maryland) under the Budapest Treaty.

Methods for incorporating a particular region of nucleic acid into a nucleic acid vector are well known in the art of molecular biology (see Sambrook, et al., supra). For example, short regions of nucleic acid (less than 200 bp) can be prepared by generating overlapping oligonucleotide fragments complementary to the nucleic acid. These oligonucleotides are then hybridized to one another, ligated and incorporated into an appropriate expression vector. Alternatively, those with skill in the art of molecular biology will be able to use the polymerase chain reaction technology to amplify a suitable nucleic acid fragment containing the regulatory gene of interest

-27-

and incorporate this fragment into the expression vector of choice (see generally Erlich, H.A. PCR Technology: Principles and Applications for DNA Amplification, 1992. W.H. Freeman and Co., New York).

5           As noted in the preceding paragraphs, the nucleic acid vector additionally contains a nucleic acid sequence encoding a recombinant protein product such that the regulatory region is operably linked to and/or regulates expression of the protein encoded by the gene. Since the  
10 regulatory region may be positioned in the nucleic acid vector as part of the nucleic acid sequence encoding protein or as a nucleic acid sequence adjacent to the nucleic acid sequence encoding protein, the assembly of the nucleic acid vector containing the regulatory region to be  
15 mutagenized may require multiple steps. Those skilled in the art of molecular biology will be able to select an initial nucleic acid vector suitable for expressing a protein product from a gene sequence in a host cell and to incorporate both the gene sequence encoding the protein  
20 product and the gene sequence encoding the regulatory region into the vector using techniques of genetic engineering well known in the art. Incorporation of both the regulatory sequence and the gene sequence encoding the protein product into the nucleic acid vector is designed  
25 such that once the nucleic acid vector is introduced into a compatible cell sample, if the regulatory sequence is functional, protein expression will be detected. Compatible cells are those cells capable of expressing a protein from the nucleic acid vector when the vector contains those  
30 regulatory elements that facilitate protein expression (i.e. origins of replication, etc.).

It is anticipated that other gene sequences may additionally be incorporated into the nucleic acid vector of this invention. For example, gene sequences coding for  
35 antibiotic resistance, or other gene sequences that

**SUBSTITUTE SHEET (RULE 26)**

-28-

facilitate cell selection may be introduced into the vector. Similarly, the nucleic acid sequence encoding protein can itself encode an antibiotic resistance gene or a portion of an antibiotic resistance gene. Other  
5 sequences that may be incorporated into the vector include but are not limited to promoters, enhancers, polyA+ adenylation sites, origins of replication (eukaryotic and/or prokaryotic), and specific restriction endonuclease cleavage sites (or arrays of sites). Exemplary nucleic  
10 acid vectors containing a regulatory region to be mutagenized and a nucleic acid sequence encoding protein are provided in Examples 1-3 as well as Figures 1-3 and 6.

**Expression of the Recombinant Protein under Control of the  
15 Regulatory Region to be Mutagenized**

In order to identify enhanced or optimized protein expression from a particular mutagenized regulatory region relative to the regulatory region to be mutagenized, some base line information should be obtained for protein  
20 produced under the control of the nonmutagenized regulatory region. This data is later compared to the level of protein expression influenced by the mutated regulatory region. This data may be obtained from the literature, or the data may be obtained first hand. As one method for  
25 obtaining the data first hand, the vector containing the nonmutagenized regulatory region is introduced into a cell sample that is able to support the expression of the gene sequence under the control of the regulatory region.

There are a variety of methods recognized in the  
30 art for introducing nucleic acid vectors into prokaryotic and eukaryotic cells. Depending on whether the cells are prokaryotic or eukaryotic, the vectors may be introduced as viral vectors, electroporated into the cells, bacterial cell walls may be permeabilized, the vectors may be  
35 introduced through standard precipitation methods or

-29-

through the use of lipophilic agents. Such methods are commercially available as kits and detailed methods are readily available in the literature. Methods for transforming *E. coli* with nucleic acid vectors are provided in Example 1.

Next, the level of protein expression from the nonmutagenized nucleic acid vector is measured in the cell sample. Proteins of interest may be concentrated from cell supernatants or cell lysates and quantitated by chromatographic methods, enzyme-linked immunoadsorbant assays (ELISA), western blot assays, radio-immunoassays, gel electrophoresis, or the like. Exemplary assays for measuring the amount of protein expressed in the periplasmic space of *E. coli* are provided in Example 1 and are detailed in Stemmer, et al. *BioTechniques* 13:114-220, 1992, hereby incorporated by reference and U.S. Patent Application No. 07/641,140, filed April 26, 1991, and U.S. Patent Application No. 07/806,154, filed December 12, 1991. Alternatively, an exemplary assay for measuring the amount of protein expressed in a eukaryotic cell is provided in Example 4. Methods to detect protein expression from a nucleic acid vector in a cell will vary depending on the gene sequence encoding protein. Those with skill in the art of molecular biology are well versed at selecting gene sequences encoding protein and matching these sequences with methods suitable for detecting protein expression derived from that gene sequence in a particular cell type. The value of protein expression obtained from cells expressing recombinant protein under the influence of the nonmutagenized regulatory region provides a level of expression that can be compared with that of cells expressing protein from vectors with mutagenized regulatory regions.

-30-

**Deletion of the Regulatory Region to be Mutagenized**

An important aspect of this invention involves the deletion of the region to be mutagenized before the pool of random oligonucleotides are incorporated into the vector. This step of the invention advantageously removes the endogenous background level of protein expression that contaminates mutagenic libraries produced by other methods. The regulatory region can be deleted from the nucleic acid vector in a variety of ways. For example, during vector construction, two identical vectors may be simultaneously assembled such that one contains and the other lacks the mutagenized regulatory region. The regulatory region can be deleted from the vector using PCR technology and an exemplary strategy is disclosed in Example 1. Similarly the regulatory region can be deleted through the use of restriction endonuclease recognition sites bracketing the regulatory region.

**Preparation of Random Oligonucleotides**

As a next step for practicing the methods of this invention, a pool of random oligonucleotides is prepared that are suitable for PCR methods and can be readily incorporated into the position in the vector previously occupied by the regulatory region. These oligonucleotides are positioned in the nucleic acid vector in place of the deleted regulatory region. As noted previously, the term "random oligonucleotides" includes biased oligonucleotides. Methods for generating random oligonucleotides on a DNA synthesizer are well known in the art. In this invention the random oligonucleotides are of a size and design useful for directing a polymerase chain reaction and for incorporating the nucleic acid vector containing the nucleic acid sequence encoding protein. That is, the oligonucleotides should contain regions of homology to the vector, they should include a restriction endonuclease site

-31-

to facilitate recircularization of the plasmid following incorporation of the vector, and they should include a region containing the random or biased nucleotides. The oligonucleotides can be of any desired length, and the  
5 length of the oligonucleotide will depend on the type of regulatory sequence and the number of random oligonucleotides that will be incorporated into the vector.

The oligonucleotides prepared in this invention contain at least one region that incorporates more than one  
10 nucleotide at a given position within the oligonucleotide. This pool of random oligonucleotides is incorporated into the vector to produce a library of clones containing random regulatory sequences. Libraries prepared from random oligonucleotides, and particularly libraries prepared using  
15 the methods of this invention wherein the regulatory region to be mutagenized is deleted, have the advantage that the selected active clones will likely be novel regulatory sequences. Because of the large theoretical number of random regulatory regions be they signal peptides, ribosome  
20 binding sites, or the like, the methods of this invention are more likely to generate novel regulatory sequences that direct increased or optimized protein expression than other methods currently available in the art.

One potential difficulty of a completely random  
25 regulatory library is that the majority of the sequences generated are inactive. This means that the libraries must be sufficiently large in order to identify a sufficient number of active clones for further study. If the goal of the mutagenesis procedure is to generate a regulatory  
30 region that significantly increases the expression of the recombinant protein, then this type of regulatory region would occur infrequently in a random library. For example, in an *E. coli* library, if the probability of finding a "super sequence" is  $10^{-9}$  there is a good chance that it  
35 would not appear in a library of  $10^8$  clones.

SUBSTITUTE SHEET (RULE 26)

-32-

**Exemplary Calculation to Estimate the Library Size  
Necessary for Identifying Novel Regulatory Sequences**

The following example is provided as one method for estimating an adequate library size that would be required to identify a novel, optimized sequence. This example is specifically addressed to the library size required to identify an optimized signal peptide in a prokaryotic cell using biased oligonucleotides. It is contemplated that those with skill in the art would be able to apply the strategy exemplified by the calculations used in this example to determine the library size required for other regulatory regions in either prokaryotes or eukaryotes.

The known requirements for a signal sequence are first incorporated into the calculation. Known signal peptides contain one or two positively charged amino acids near the N-terminus, a stretch of 8-12 hydrophobic amino acids and a small amino acid such as alanine, glycine or valine positioned at the C-terminus. For consensus regions in prokaryotic signal sequences see MacIntyre, et al. *Mol. Gen. Genet.* 221:466-474, 1990; Gennity, et al. *J. Bioenergetics and Biomembranes* 22:233-269, 1990; Laforet, et al. *J. Biol. Chem.* 266:1326-1334, 1991 and Yamane, et al. *J. Biol. Chem.* 263:19690-19696, 1988 and for consensus regions and exemplary eukaryotic signal sequences see Von Heijne, et al. *Eur. J. Biochem.* 133:17-21, 1983.

The fraction of active leader peptides in this example is  $(f) = p_1^+ \cdot p_2^+ \cdot p_3^+ \dots p_n$ , where  $p_1$ ,  $p_2$ ,  $p_3 \dots p_n$  represent the probability of a given amino acid or amino acids, occurring in a given position. In a hypothetical library containing a 22 amino acid signal peptide, the peptide can be diagrammed as  $[M][+][+][n]_8[S]$  where each  $[]$  represents an amino acid,  $[M]$  represents methionine,  $[+]$  represents a positively charged amino acid,  $[n]$  represents any hydrophobic amino acid and  $[S]$  represents a tiny amino

-33-

acid. The fraction of active leader peptides (f) =  
(1)(.15)(.15)(.66)<sup>18</sup>(.2) = 7.3 x 10<sup>-6</sup> Pugsley, supra, p. 47,  
Sambrook, supra, Appendix p. 6. Thus a signal peptide of  
this particular configuration would occur once in 300,000  
5 random clones. This value reflects the incidence of an  
active signal peptide within a library and not the  
frequency of an optimized signal peptide. One would expect  
that the frequency of an optimized signal peptide occurring  
in a random library to be less; therefore the library size  
10 would need to be larger.

Several assumptions were made in the above  
calculation. First, it was assumed that two positive  
charges are required at the N-terminus. In fact, there are  
known signal peptides that carry only a single positive  
15 charge at their N-terminus. If only a single positive  
charge is required, active leader peptides will occur at a  
higher frequency. Similarly, for this calculation, the  
hydrophobic region was limited to only hydrophobic amino  
acids. There are some cases where amino acids such as gln  
20 or asn or even acidic or basic residues occur in this  
region. Thus, the calculation of (0.66)<sup>18</sup> may be too  
restrictive. If 2 amino acids in the hydrophobic stretch  
are permitted to be any of the 20 amino acids, then we  
arrive at a value of (0.66)<sup>16</sup>. Finally, the impact of the  
25 degeneracy of the code was ignored in these calculations.  
Although the last amino acids in known signal peptides are  
usually glycine or alanine, these amino acids are generally  
believed to be required for signal peptide cleavage. Since  
signal peptide cleavage is not required for translocation  
30 and  $\beta$ -lactamase activity, that requirement can be  
eliminated. If these less restrictive requirements are  
used, the frequency of a leader peptide occurring in a  
random library is 0.02% or 1 in 5000 random sequences.  
These types of calculation are important for estimating the

-34-

required library size needed to identify a novel, optimized sequence.

In practical terms, a biased library is difficult to construct due to the constraints on the usage of the genetic code. For example, if one tries to use oligonucleotide doping to produce either a lysine or arginine in a given position, the codon for glutamine will also occur. Similarly, due to the degeneracy of the genetic code, it is impossible to constrain an amino acid to solely a hydrophobic amino acid using single base oligonucleotide synthesis from variable base mixtures. In some cases it is possible to use doping to achieve the desired amino acid mixture. For example, a 50% occurrence of alanine or glycine will occur if the sequence 5' G(G/C)X is used. The successive addition of pre-made trimers can obviate this problem. Precise ratios of desired amino acids can be designed into the library through use of premixed trinucleotide codon triplets during DNA synthesis (Sondek, et al. *Proc. Natl. Acad. Sci. (USA)* 89:3581-3585, 1985). Given current DNA synthesizer technology this would be a time consuming and expensive undertaking.

As another method for generating a biased library the oligonucleotide column swapping DNA synthesis technique developed by Glaser, et al. is used (*J. Immunol.* 149:3903-3913). This method produces a 50% chance of a mutation at any desired amino acid. This is a codon based mutagenesis approach that allows the formation of large diverse libraries. The column-swapping manipulation causes one-half of each synthesis step to contain random codons, while the other half contains wild-type codons, at each position to be mutated. In this way, one can mutate large stretches of DNA without a high frequency of stop codons, as only one of the three possible stop codons can occur, and this only at a frequency of 1 in 32. The use of a suppressor tRNA for this stop codon can reduce the frequency to zero.

SUBSTITUTE SHEET (RULE 26)

-35-

However, the preparation of biased oligonucleotides to identify optimized regulatory sequences is not an absolute requirement for the methods of this invention. Optimized sequences can be identified from  
5 clones incorporating random oligonucleotides containing no chemical bias. In a preferred embodiment, illustrated in Example 1, the invention demonstrates that even with no preconceived notions about the specific requirements for a regulatory region, the bacterial host will naturally select  
10 those regulatory regions that it finds functional in expressing the protein of interest.

#### Incorporation of the Oligonucleotides into the Nucleic Acid Vector

15 As a next step for practicing the methods of this invention, the mutagenized oligonucleotides are introduced using a polymerase chain reaction into the position in the vector previously occupied by the deleted regulatory region. There are a variety of polymerase chain reaction  
20 methods known in the art that are suitable to the methods of this invention. The examples below use the enzymatic inverse polymerase chain reaction (see Stemmer, et al. *BioTechniques* 13:114-220, 1992, hereby incorporated by reference and U.S. Patent Application No. 07/641,140, filed  
25 April 26, 1991, and U.S. Patent Application No. 07/806,154, filed December 12, 1991) as a preferred PCR strategy. EIPCR advantageously allows targeted mutagenesis of a given DNA sequence and rapid selection of active clones from a large library. Other PCR techniques suitable for use in  
30 the methods of this invention include, but are not limited to, inverse PCR (Hemsley, et al. *Nucl. Acids Res.* 17:6545-6551, 1989), recombinant circle PCR mutagenesis (Jones, et al. *BioTechniques* 8:178-183, 1990), recombination PCR (Jones, et al., *BioTechniques* 10:62-66, 1991) and standard

-36-

PCR techniques such as those described by Saiki, et al. (Science 239:487-491, 1988).

Briefly, the enzymatic inverse polymerase chain reaction (EIPCR) involves the use of oligonucleotides containing a class IIS restriction endonuclease recognition site and the associated downstream class IIS cleavage site. The EIPCR technique uses a small circular vector and the PCR primers are designed to overlap at their class IIS cleavage sites. In one embodiment, one oligonucleotide of the oligonucleotide pair used in the EIPCR reaction contains in a first 5' to 3' orientation a 5' tail, a class IIS restriction endonuclease recognition site, a class IIS cleavage site, the region containing the mutation and a region complementary to the nucleic acid vector. The other oligonucleotide of the oligonucleotide pair comprises in a second 5' to 3' orientation; a 5' tail, a class IIS restriction endonuclease recognition site, a class IIS cleavage site and a region complementary to the nucleic acid vector. In another preferred embodiment, both oligonucleotides of the oligonucleotide pair used in the EIPCR reaction contain the first 5' to 3' orientation. Further details of primer design applicable to the mutagenesis application of this invention are provided in Examples 1-3.

The amplified linear vectors are digested with the appropriate restriction endonuclease and recircularized to produce a pool of mutagenized vectors. The pool of mutagenized vectors is purified and aliquots are introduced into a cell sample using any of the methods disclosed in association with the discussion relating to cell transformation or transfection discussed above. Example 1 uses electroporation to introduce the mutagenized vectors into the cell sample. Preferably the cell samples used for both the mutagenized and nonmutagenized vector transformations or transfections are derived from the same

-37-

cell type. In this way, protein expression from cells containing the nucleic acid vector having the nonmutagenized regulatory region can be directly compared to expression from cells containing the nucleic acid vector with the mutagenized regulatory region.

#### Assaying for the Expression of the Recombinant Protein from the Mutagenized Vectors

Following introduction of the mutagenized vectors into the cell sample, the cell sample is assayed for the expression of the recombinant protein, that is, the protein associated with the mutagenized regulatory region. As noted in the section entitled "Expression of the Recombinant Protein under Control of the Regulatory Region to be Mutagenized", the particular assay for assessing the level of protein expression will depend on the particular protein being expressed as well as on the cell type. Exemplary methods for assessing protein expression in the periplasmic space of prokaryotes are provided in Example 1. For enhanced protein expression, cells are selected that exhibit protein expression from the nucleic acid vector at a level greater than or equal to the level of protein expression observed in the cell sample containing nonmutagenized vector. For optimized protein expression, other characteristics, in addition to the level of protein synthesis, may be important for clone selection. These include cell growth characteristics, the location of protein expression or the like.

Referring to Example 1 below, the processes of transcription, translation, translocation and signal peptide cleavage are parameters that are affected when the signal sequence is mutated. The use of a random signal sequence allows for selection of clones that are optimized for all of these parameters using the internal mechanisms of the host cell. For example, a random signal sequence

-38-

that generates a nonfunctional secondary structure in RNA would preclude efficient translation and positive clones containing this sequence would not be identified. Assays for recombinant protein expression from the mutagenized  
5 vectors permits the cumulative effect of the mutated sequence on protein expression to be quantitated. In addition, these parameters are optimized with respect to the individual requirements of the product proteins being expressed by the regulatory regions such as glycosylation  
10 requirements, folding requirements for functionality, and the like.

As a final step in the methods of this invention, the nucleic acid vector is isolated from cells derived from the cell sample that contain mutagenized vector and exhibit  
15 enhanced or optimized expression of the recombinant protein. The purified vector is subjected to DNA sequence analysis to determine the nucleic acid sequence of the mutagenized regulatory region corresponding to the random oligonucleotide. DNA sequencing techniques are well known  
20 in the field of molecular biology, thus no further disclosure of sequencing techniques is required.

Identification of optimized regulatory sequences is a particularly important step in the method of optimizing protein expression from any number of vectors.  
25 Thus, this invention provides a method for both optimizing regulatory regions associated with protein expression and for identifying the optimized sequences of these regions. Those with skill in the art of molecular biology will be able to incorporate these optimized sequences into other  
30 constructs thereby optimizing expression of any number of proteins from a variety of vectors introduced into a variety of cell types.

SUBSTITUTE SHEET (RULE 26)

-39-

### Incorporation of Selectable Markers into the Nucleic Acid Vector

Rather than improving upon existing regulatory regions, this invention creates novel regulatory sequences, including novel signal peptides. As noted in the section entitled "Exemplary Calculation to Estimate the Library Size Necessary for Identifying Novel Regulatory Sequences", one would expect that a library of 21 random amino acid sequences would contain a relatively low percentage of functional secreting clones. However, by attaching the random regulatory region to a selectable marker, particularly a marker which requires secretion for activity, the library can be restricted to only clones containing functional signal peptides. Selectable marker proteins, particularly those in prokaryotes require transport to the periplasmic space to be active, have the advantage that when a regulatory region is operably linked to the amino-terminus of a heterologous protein containing the C-terminus of a periplasmic protein, this will result in transport to the periplasmic space and the selectable marker is able to confer a subsequent advantage to cell growth under certain conditions. The in-frame fusion of a heterologous protein to a selectable marker provides a rapid means to select only active regulatory sequences. Furthermore, if the selectable marker is an antibiotic resistant gene, then the use of increasing concentrations of antibiotic allows for selection of clones containing the most efficient regulatory sequences.

The enzyme  $\beta$ -lactamase represents one such selectable marker because it is active only if translocated from the cytoplasm to the periplasm (Pluckthun, et al. *J. Biol. Chem.* 262(9): 3951-3957, 1987 and Francisco, et al. supra). While Example 1 uses a regulatory region operably linked to a eukaryotic protein fused to the C-terminus of  $\beta$ -lactamase, it also is contemplated that the regulatory

SUBSTITUTE SHEET (RULE 26)

-40-

region can be operably linked to a full-length selectable marker such as  $\beta$ -lactamase (see Example 3). However, the creation of novel regulatory regions by assessing protein expression in the periplasmic space is not limited to

5  $\beta$ -lactamase fusion proteins. Other selectable marker proteins can also be used. As an alternative, aminoglycoside 3'-phosphotransferase II can be used as a selectable marker in the same way. This enzyme is present in bacteria that are resistant to streptomycin, neomycin or

10 kanamycin. The enzyme renders the antibiotic inactive as it enters the cell through the periplasm. Other periplasmic enzymes include Streptomycin adenylating enzyme and alkyl sulfhydrolase.

Other selectable markers contemplated for use in

15 the methods of this invention include binding proteins for various carbon sources. Proper secretion of this class of proteins leads to their presence on the outer membrane or periplasm, enabling the cell to internalize a particular sugar or amino acid. For example, maltose binding protein,

20 when properly secreted, allows bacteria to use maltose as the sole carbon source. If the maltose binding protein is not secreted, cells cannot grow on media containing maltose as the sole carbon source. In some cases, it is necessary to use the appropriate auxotrophic strain of *E. coli*. For

25 example a his-negative strain requires histidine in the media. Other selectable proteins that may be useful as fusion proteins include but are not limited to arabinose binding protein, arginine binding protein, cystine and diaminopimelic binding protein, galactose-glucose binding

30 protein, glutamate-aspartate binding protein, glutamine binding protein, histidine binding protein, leucine specific binding protein, lysine-arginine-ornithine binding protein, phosphate binding protein, ribose binding protein, sulfate binding protein, thiamine binding protein and

35 xylose binding protein.

-41-

One can also use nonselectable markers. These are proteins that do not necessarily give the cell any physiological advantage. Their presence in the periplasm is detected based on their interaction with specific  
5 ligands. With the use of these markers it is possible to detect the periplasmic protein by probing with the appropriate ligand. Examples of these markers include but are not limited to hapten binding antibodies, cytochrome c, nitrite reductase, phosphoglucomutase and phosphoglucose  
10 isomerase. Therefore, while Example 1 employs a random sequence library linked to a fusion protein of a single chain antibody with an antibody resistance gene, the regulatory region could also be linked solely to a sequence encoding an antibody sequence and positive clones selected  
15 by screening with labelled antigen. For identification of intact colonies expressing antibody sequences in the periplasmic space, the antigen is preferably a small hapten capable of diffusing into the periplasmic space.

As noted above, any regulatory region affecting  
20 the level of protein synthesis is contemplated as a potential target for the optimization strategies of this invention. Two specific prokaryotic examples are provided below. These examples illustrate the methods of this invention as they apply to optimizing signal sequences  
25 (Example 1) and ribosome binding sites (Example 3). The methods of this invention readily apply to mutagenesis strategies directed to both these and other regulatory regions in prokaryotic and eukaryotic cells.

As a particularly preferred strategy for  
30 generating optimized signal sequences encoding signal peptide, Example 1 details a method for creating novel signal sequences for the expression of eukaryotic protein in *E. coli*. In addition to creating novel signal sequences, this method is also useful for optimizing the

SUBSTITUTE SHEET (RULE 26)

-42-

expression of a particular protein from a given construct in a given cell type.

The strategy detailed in Example 1 combines the mutagenesis strategies of this invention with antibiotic selection to improve secretion of Fv antibody molecules in *E. coli*. A fusion protein consisting of an N-terminal, hapten-binding single chain antibody (CHA 255) and the C-terminal of  $\beta$ -lactamase was constructed and expressed in active form in the periplasmic space of *E. coli* using the Omp A signal peptide. The fusion of  $\beta$ -lactamase with a second protein to study protein transport was reported by Broome-Smith, et al. *Mol. Micro.* 4(10) 1637-1644, 1990. For studies linking Omp A with a prokaryotic protein as a fusion protein to  $\beta$ -lactamase see Francisco, et al. *Proc. Natl. Acad. Sci. USA* 89:2713-2717, 1992 and Meyer et al. disclosing a method for combining an Fab' antibody conjugate to  $\beta$ -lactamase, *Bioconj. Chem.* 3:42-48, 1992). For site-directed mutagenesis of a eukaryotic signal sequence associated with a eukaryotic protein fused to the C-terminus of  $\beta$ -lactamase for prokaryotic expression see Stahl, et al. (*Gene* 71:147-156, 1988)

Unlike previous mutagenesis strategies, the regulatory sequence of interest, here the Omp A signal peptide, was deleted and enzymatic inverse PCR was used to construct a library containing random signal peptides at the N-terminus of the fusion protein. None of the references cited above report the deletion of the regulatory region to be mutagenized followed by the isolation of vector containing that deletion to create large libraries of regulatory mutants uncontaminated with false positive clones. Functional signal peptides were selected by plating the clones on media containing ampicillin. Clones were identified that had  $\beta$ -lactamase activity comparable to or greater than that of the fusion protein containing the Omp A signal peptide. To ensure

SUBSTITUTE SHEET (RULE 26)

-43-

retention of antibody activity, colony filter lifts were screened with a radioactively labeled hapten, <sup>111</sup>In-EOTUBE, recognized by the CHA 255 antibody.

Example 1 provides a detailed method for  
5 generating novel signal sequences that are 22 amino acids in length. However, optimized signal sequences, like optimized regulatory regions can be selected based on the lengths of known regulatory regions, or, alternatively, regulatory regions of any desired length can be selected  
10 for optimization. Therefore, steps similar to those outlined in Example 1 can be performed to generate optimized signal sequences, or other regulatory regions of any length. Example 2 provides methods for creating random leader peptides of 12, 16, or 20 amino acids in length.

15 It is further contemplated that the methods of this invention can be used to create novel regulatory sequences for expression of either heterologous or homologous protein. For example, Example 3 uses a native *E. coli* protein to identify novel ribosome binding site  
20 sequences. Once optimized ribosome binding site sequences are identified they can be incorporated into other vectors to study the effect of these sequences on heterologous protein expression. In this example the steps outlined above are followed for creating a library of random  
25 regulatory sequences operably linked to protein except that here the initial nucleic acid vector construct includes the Omp A ribosome binding site linked to the Omp A signal peptide located at the amino terminus of the  $\beta$ -lactamase protein. The Omp A ribosome binding site sequence was  
30 deleted by PCR and the resulting vector was used to incorporate a library of random oligonucleotides. These mutagenized vectors were introduced into *E. coli* and the level of protein expression from the resulting clones was assessed using a PADAC assay. Three clones were identified  
35 that expressed elevated levels of protein relative to the

SUBSTITUTE SHEET (RULE 26)

-44-

nonmutated ribosome binding site sequence. These clones were selected and the nucleic acid sequence of the mutated ribosome binding site was determined. Details of this experimental protocol are provided in Example 3.

5 In yet another preferred embodiment of this invention, a method is provided for creating and identifying novel regulatory regions in eukaryotes. Example 4 details a method for producing signal peptide libraries in mammalian cells for improved protein expression. Known  
10 eukaryotic signal sequences are disclosed and reviewed by Von Heigne, G., *Eur. J. Biochem.* 133:17-21, 1983. Figure illustrates an exemplary vector contemplated for use in this invention. The vector uses the Immunoglobulin heavy  
15 chain promoter to initiate transcription of the desired gene inserted downstream from the promoter. The gene to be expressed along with the regulatory region to be mutagenized is inserted into the vector. The regulatory region is deleted and a random or biased regulatory  
20 sequence is inserted into the vector. Example 4 teaches the mutagenesis of the signal sequence. Following the incorporation of the random signal sequence into the eukaryotic vector, the vector is introduced into suitable eukaryotic cells that are capable of expressing protein from the expression vector. Example 4 employs the vector  
25 of Figure 7 in SP 2/0 cells. Vectors can be introduced into cells using those methods known in the art. Example 4 employs electroporation to mediate transfection. However, those with skill in the art will be readily able to select an appropriate method for introducing the vector into the  
30 eukaryotic cells. Positive cells are preferably selected by growing the cells in neomycin. In Example 4, the relative level of protein expression is assayed using a Western dot blot. Positive transfected cells can be further quantitated for protein expression by ELISA as  
35 compared with cells expressing protein under the control of

SUBSTITUTE SHEET (RULE 26)

-45-

the native regulatory sequence. Cells expressing equal or greater levels of protein expression as compared with cells containing the native regulatory sequence are subjected to DNA sequencing.

5 In another preferred embodiment of this invention, the methods of the invention are used to generate targeting sequences to target proteins to particular cellular locations. The terms "protein  
10 targeting sequence" and "protein trafficking sequence" are used interchangeably in this disclosure. Both prokaryotic and eukaryotic systems studied to date indicate that sequences associated with protein targeting can occur in a variety of locations within a protein sequence. For  
15 example, in bacteria, proteins with N-terminal signal sequences are secreted into the periplasmic space or are inserted into the outer membrane using a secretion apparatus defined by a series of sec genes (Bassford, et al., *Cell* 65:367, 1991). Other proteins lacking signal  
20 sequences are also secreted into the extracellular space in *E. coli* and some of these proteins contain secretion signals located at the C-terminus of the protein (Delepelaire, P. et al, *J. Biol. Chem.* 265:17118, 1990). These secretory proteins lack a classical hydrophobic  
25 signal peptide. Instead, the proteins contain a consensus region of 200 amino acids containing highly conserved stretches of amino acids that are postulated to confer secretory activity.

Similarly, targeting sequences are located within a variety of locations in eukaryotic proteins. For  
30 example, interleukin proteins lack hydrophobic leader peptides and their targeting sequences are believed to be located in the C-terminal portion of the protein precursor (Marck, et al., *Nature*, 315: 641, 1985). *E. coli* expression of Interleukin-1 results in the translocation of  
35 the protein to the periplasmic space, while expression in

**SUBSTITUTE SHEET (RULE 26)**

-46-

yeast results in expression of the protein in the cytosol. Particular protein sequences also are implicated in mitochondrial targeting (see Zara, et al., *J. Biol. Chem.* 267:12077-12081; Adrian, et al., *Mol. Cell. Biology* 6:626-634, 1986 and Horwich, et al, *Cell* 44:451-459). Similarly in both prokaryotes and eukaryotes, post-translational modifications, such as glycosylations (eukaryotes), the addition of lipids, or the like, have been shown to be important in protein translocation and targeting within a cell. It is generally believed that specific regions within the protein sequence are important determinants for post-translational modifications within a cell. Kuchler, et al, *Endocrine Reviews* 13:499-514, 1992 observed that the C-terminal modification of a chimeric IL-1<sub>γ</sub>:bacterial α-factor protein resulted in exclusive plasma membrane association. Others have demonstrated that a combination of post-translational modifications of the C-terminal region of a heterologous protein with the presence of a polybasic domain, positioned near the C terminus, result in the targeting of heterologous protein to the plasma membrane in MDCK cells (Hancock, et al. *EMBO J.* 10:4033, 1991). In another series of experiments, Bakke, et al. (*Cell* 63:707-716, 1990) identified the region of a protein and the particular sequence responsible for sorting a plasma membrane protein from the endosomal vesicles to the plasma membrane. For a review of the location and sequences of known eukaryotic sorting sequences see Schwartz, *Ann. Rev. Immunol.* 8:195-229, 1990.

The ability to target protein to a particular cell location is an important goal in gene therapy strategies. Advantageously, the method described in the claims and below and detailed in Example 5, permits one with skill in the art to introduce a particular nucleic acid sequence encoding protein into a cell and to select cells expressing protein in the desired cell location. A

-47-

comparison of the native protein sequence with the mutated protein sequence that targets the protein to the desired cell location permits the identification of the novel sequence that directs protein targeting.

5           As a first step for practicing the trafficking or targeting method of this invention, a particular region of nucleic acid containing a protein trafficking/targeting signal or a postulated protein trafficking/targeting signal is identified. As noted, trafficking signals can be  
10 positioned in a variety of locations within a protein sequence. They have been identified with the leader sequence as well as within the C-terminal portion of a protein. Those with skill in the art of molecular biology will be able to search the literature for recognized  
15 regions of eukaryotic or prokaryotic protein sequences that are postulated to affect protein trafficking and/or targeting. Equivalent regions within the nucleic acid sequence are then identified.

          Next, a nucleic acid vector is prepared that  
20 contains the nucleic acid sequence encoding the protein trafficking signal to be mutagenized and a nucleic acid sequence encoding at least one protein. The nucleic acid vector may be suitable for directing eukaryotic or prokaryotic expression of the protein, and those with skill  
25 in the art of molecular biology will be readily able to select and construct an expression vector suitable for this invention. Once the vector is prepared, the nucleic acid sequence containing the protein trafficking signal is deleted from the vector.

30           As a next step for practicing the methods of this invention, a pool of random oligonucleotides are produced that are suitable for use in a polymerase chain reaction. These oligonucleotides are incorporated into the position in the vector previously occupied by the protein  
35 trafficking signal to generate a pool of mutagenized

-48-

vectors. Suggested methods for the steps of this invention are discussed supra as they relate to methods for creating nucleic acid regulatory sequences and specific examples of the suggested methods are provided in the Examples below.

5           The mutagenized vectors are introduced into a cell sample using electroporation, membrane permeabilization,  $\text{CaCl}_2$  precipitation, viral vectors, or the like. Those with skill in the art will use their preferred methods for introducing nucleic acid vectors into a cell  
10 sample.

          As a next step for practicing the methods of this invention, the cells are assayed to determine the location of the protein encoded by the nucleic acid vector in the cell sample. Methods to determine the localization of a  
15 protein within a cell are known in the art. For example, if antibodies are available that react with the protein encoded by the nucleic acid vector, then these antibodies can be labelled with a fluorescent marker, colloidal gold, or the like. Cell samples expressing the protein can be  
20 membrane permeabilized for fluorescent antibody analysis or cells can be fixed and sectioned for protein localization studies using electron microscopy in the presence of colloidal gold. In addition, cell samples can be  
fractionated and individual organelles separated and  
25 individually tested for the presence of the protein. Cells expressing the protein in the desired cell location are identified by an assay and the nucleic acid sequence of the novel protein trafficking/targeting signal is determined and compared to the native sequence.

30           Particular embodiments of the invention will be discussed in detail in the following examples and reference will be made to possible variations within the scope of the invention. There are a variety of alternative techniques and procedures available to those of skill in the art which

-49-

would similarly permit one to successfully perform the intended invention.

#### Example 1

##### 5 Creation of Novel Signal Sequences in *E. Coli*

This example illustrates a method for producing novel signal sequences for optimized expression of eukaryotic or prokaryotic protein in prokaryotic cells.

The SCCHA225  $\beta$ -lactamase fusion protein was  
10 cloned as follows: SCCHA225 was amplified from the construct pUCHAsc2 using a 5'-terminal primer (B444, designated SEQ ID NO: 1, see Figure 1) containing an Xba I restriction site and a 3' primer (B456, designated SEQ ID NO: 2) containing DNA coding for the (gly<sub>4</sub>ser)<sub>2</sub> sequence and  
15 a Class IIS restriction endonuclease restriction site, Bsa I. The designation "2" in the name pUCHAsc2, indicates that the ampR gene was a mutation eliminating the BSA1 site, which was accomplished by a single round of EIPCR using primers 939 (SEQ ID NO: 3) and 940 (SEQ ID NO: 4) as  
20 described in Example 2).  $\beta$ -lactamase was amplified from pUCHAsc2, which contains a modified copy of  $\beta$ -lactamase lacking the Bsa I restriction site, but containing the native amino acid sequence. A 5'-terminal primer (B457, designated as SEQ ID NO: 5) containing a Bsa I restriction  
25 site and nucleotides coding for (gly<sub>4</sub>ser)<sub>2</sub> were used. The 3'-overlap primer (B447, designated as SEQ ID NO: 6) introduced 3 consecutive stop codons and a Bam HI restriction site at the 3' end of the  $\beta$ -lactamase gene. The use of overlapping primers, B456 and B457, created a 10  
30 amino acid (gly<sub>4</sub>ser)<sub>2</sub> sequence between the single chain antibody and  $\beta$ -lactamase domains (Figure 1). The resulting PCR products were processed with Bsa I, ligated and the ligation products were used as template for a second PCR reaction using the external primers B444 and B447. Ligation  
35 conditions are provided below. The 1.6 kb PCR product was

-50-

then gel purified and subcloned into the expression plasmid pCCHAsc1 cut with Xba I and Bam HI. pCCHAsc1 is a pUC based plasmid that constitutively expresses a chloramphenicol resistance gene and contains an expression cassette having  
5 a Lac promotor driving expression of the CHA 255 single chain antibody through the use of an Omp A derived ribosome binding site and signal sequence. The Omp A sequence was created from primers using the sequence provided in the literature (Movva, et al. *J. Mol. Biol.* 147:317-328, 1980).  
10 The resulting plasmid pCCHAsc1-ampRV2 (see Figure 1) was verified by restriction digestion and DNA sequencing. Double stranded dideoxy sequencing was performed on a Dupont Genesis 2000, using the DuPont Genesis 2000 sequencing kit according to the manufacturer's  
15 instructions. Post gel processing was done with the Base Caller 5.0 program (DuPont, Boston, MA).

Oligonucleotides used to produce the fusion protein construct are shown in Figures 1 and 2. Oligonucleotides were synthesized in an Eppendorf Synostat  
20 D automated DNA synthesizer (Madison, Wisconsin). Oligonucleotides used for library EIPCR were synthesized with the 5'-trityl group on and were purified with a Nensorb Prep column (New England Nuclear, Tozer, MA) according to the manufacturer's instructions.  
25 Oligonucleotides containing random nucleotide incorporation were prepared by selecting a mixed base option according to the software provided by the manufacturer.

Standard ligations contained 0.1-1.0  $\mu$ g DNA, 1x ligation buffer and 1-2  $\mu$ l (400,000 units) of T4 DNA ligase  
30 (New England Biolabs, Tozer, MA) in a 20  $\mu$ l volume. The ligation reactions were incubated at room temperature for one hour. Mass estimates for chemically synthesized oligonucleotides were obtained spectrophotometrically. In all other cases, DNA mass estimates were made by visual  
35 comparison of samples run on agarose gels with a 1 kb

-51-

ladder DNA standard (GIBCO/BRL, Gaithersburg, MD) loaded to approximate 100 ng DNA per band.

The construct pCCHAScl-ampRv-2 was introduced into *E. coli* DH-10B (GIBCO-BRL, Gaithersburg, MD). Bacteria were propagated at 30°C in Terrific Broth medium (Sambrook, et al. 1989. supra) or on agar plates containing 34 µg/ml chloramphenicol and 100 µg/ml ampicillin. Positive clones were selected from the ampicillin plates for further study. Assays to monitor the level of protein expression in the bacteria are described below. The completed vector provides the template for EIPCR-based construction of the signal peptide library.

For EIPCR, the PCR reactions contained 0.2 ng the template, 0.5 µM of each primer, 1X Taq buffer (Perkin-Elmer Cetus, Norwalk, CT), 200 µM of each dNTP, 1.61 mM MgCl<sub>2</sub>, and 5 units of Taq DNA Polymerase (AmpliTaq™, Perkin-Elmer Cetus) in a total volume of 100 µl. The addition of MgCl<sub>2</sub> beyond 1.61 mM up to 3.11 mM is often used as a variable that when modified impacts the yield of DNA. Templates were amplified in a Perkin Elmer Thermocycler (Norwalk, CT) using a "hot start" according to manufacturer's instructions (3 min. at 94°C) followed by 30 cycles of 94°C (1 min.) denaturation, 50°C (1 min.) annealing and 72°C (2 min.) extension. A 5 second auto extension was used. Extension was completed with an additional 10 minute incubation at 72°C.

The EIPCR template DNA is similar in size to the desired EIPCR product, hence they could copurify by agarose gel electrophoresis during further manipulations. While the circular template is present at a much lower concentration than the EIPCR product, it will electroporate at high efficiency in bacteria thereby resulting in the contamination of the library with wild-type clones. The data presented in Table 1, below, demonstrate the problem of wild type clone contamination in library mutagenesis.

SUBSTITUTE SHEET (RULE 26)

-52-

When pCCHAScl-ampR was used as a template, only the wild-type sequence was recovered from the library under selective conditions.

To render the expression construct nonfunctional and eliminate the appearance of functional template in the EIPCR product, the entire Omp A signal sequence was deleted from pCCHAScl-ampRV-2 in a single EIPCR reaction (see Figure 2A). Two primers, B593 (SEQ ID NO: 7) and B594 (SEQ ID NO: 8), were used in an enzymatic inverse polymerase chain reaction to remove the signal sequence. The oligonucleotides hybridized to opposite strands of the plasmid and in a PCR reaction yielded a linear DNA that is slightly smaller than the original template. This linear DNA contains at its termini a 14 bp extension containing the recognition site for the class IIs restriction site Bsa I. The terminal extensions of DNA were removed by digestion with Bsa I, yielding compatible cohesive ends that efficiently undergo an intramolecular ligation. The resulting plasmid, pCCHAScl-ampR\_SP, is a 3.2 kB pUC-derived plasmid which contains a single chain CHA225- $\beta$ -lactamase fusion protein absent the Omp A signal peptide, and this vector confers resistance to chloramphenicol. Removal of the signal sequence resulted in ampicillin sensitivity. Thus, any pCCHAScl-ampR\_SP clones that appeared in the final EIPCR library, were eliminated during selective expression to yield the desired mutagenized clones (see Table 1, below).

The random oligonucleotides, created as described above, were incorporated into pCCHAScl-ampRV2\_SP using a polymerase chain reaction. Figure 2b illustrates the use of EIPCR to create a random signal peptide library. Divergent oligonucleotide primers were used to amplify the DNA and to incorporate mutant sequences. The forward primer, B524 (SEQ ID NO: 9), annealed to the DNA corresponding to the N-terminal region of the scCHA

SUBSTITUTE SHEET (RULE 26)

-53-

antibody. The reverse primer, B525 (SEQ ID NO: 10), incorporated the initiator codon and 21 random codon triplets. The reverse primer's annealing location is illustrated in Figure 2b. Both the forward and reverse  
5 primers contained the Bsa I restriction site near their termini. The Bsa I restriction site is present at both ends of the linear EIPCR product and was removed by digestion, yielding compatible cohesive ends. The linear PCR product was then ligated and electroporated into *E.*  
10 *coli*.

The infrequent occurrence of an active signal peptide in a library of random signal peptides necessitated several considerations for library construction. First, to ensure a large enough library size and complexity we used a  
15 higher than normal amount of template DNA in our EIPCR reactions. Typically EIPCR reactions contain only 0.5 ng of template DNA per 100  $\mu$ l reaction (Stemmer, et al. supra). Increasing the amount of template to 25-50 ng per 100  $\mu$ l reaction resulted in more EIPCR product and a larger  
20 library size. As an example, when pCCHAScl-ampRV2 was used as a template, 5 ng of plasmid yielded only 0.25  $\mu$ g of PCR product. When 50 ng of template was used, 5  $\mu$ g of PCR product was obtained. It is contemplated that the activity of different primer-template combinations will be optimized  
25 within a range of template concentrations from about 0.1 to 50 ng of template per 100  $\mu$ l reaction. Those with skill in the art of molecular biology can readily determine the optimal template concentrations as well as adjust the concentration of  $MgCl_2$  to maximize PCR reaction product.  
30 Advantageously, the deletion of the critical regulatory sequence from the template prior to the incorporation of random sequence enables one to use greater amounts of template than those typically used in standard PCR reactions without concern for amplifying nonmutated  
35 template. EIPCR reaction conditions for the incorporation

SUBSTITUTE SHEET (RULE 26)

-54-

of random oligonucleotides were the same as for the cloning reactions.

Table 1

5	Library	Template	<u>CFU/<math>\mu</math>g DNA</u>		% Wildtype (n)
	Template	DNA/100 $\mu$ l	CAM	CAM/AMP	CAM/AMP
10	pcCHAscl-ampR	20ng	20.4x10 <sup>6</sup>	4511	100% (8)*
	pcCHAscl-	50ng	3.9	1904	0 (10)**
15	AmpRASP				

CAM (chloramphenicol) = 34  $\mu$ g/ml

AMP (ampicillin) = 50  $\mu$ g/ml

n=number of samples evaluated

20 \*=determined by restriction mapping

\*\*\*=determined by DNA sequencing

The 100  $\mu$ l PCR mixture was extracted with an equal volume of phenol-chloroform-isoamyl alcohol (24:23:1) and was then precipitated using 2 volumes of ethanol. The protruding termini of the PCR end products were filled in and digested according to the method of Stemmer (*Biotechniques*, 1992 supra). Briefly, the pellet was resuspended in 70  $\mu$ l H<sub>2</sub>O and 10  $\mu$ l of 10X Klenow buffer (NEB, Tozer, MA), 10  $\mu$ l dNTP mix (2.5 mM each dNTP), 5  $\mu$ l of DNA Polymerase I (large fragment: Klenow), and 5  $\mu$ l of T4 DNA polymerase were added. The reaction was incubated at 37°C for one hour followed by phenol extraction and ethanol precipitation as disclosed above. Blunt-ended EIPCR fragments were then exhaustively digested in a 100  $\mu$ l reaction containing 50 units of Bsa I at 60°C for four hours. The Bsa I digested fragment was purified away from both the PCR primers and the short terminal DNA fragment by

SUBSTITUTE SHEET (RULE 26)

-55-

centrifugation through a Biospin 6 column (BioRad, Richmond, CA) according to the manufacturer's protocol. The purified Bsa I digested fragment was then phenol extracted and ethanol precipitated as described earlier and  
5 ligated under standard conditions with the exception that the fragments were ligated at higher concentrations of T4 DNA ligase (106 units/ $\mu$ l) and incubation was at 12°C overnight.

Ligated or control DNA (pUC 19, GIBCO/BRL) was  
10 precipitated in 2 volumes of ethanol and resuspended in 20  $\mu$ l of TE (10mM Tris, pH 7.4, 0.1 mM EDTA, pH 8.0). 10  $\mu$ g of yeast tRNA (GIBCO/BRL) was added as a carrier and did not affect electroporation efficiency. The DNA was electroporated in 1-5  $\mu$ l aliquots into electrocompetent  
15 DH10-B MAX cells (BRL, Bethesda, MD) according to the manufacturer's instructions. pUC19 DNA (0.01 ng) was electroporated to monitor the electroporation efficiency.

Under conditions to select for the plasmid marker (34  $\mu$ g/ml chloramphenicol), a library of  $3.9 \times 10^6$  colony  
20 forming units was obtained from approximately 1  $\mu$ g of ligated DNA (Table 1). Under dual selective conditions (34  $\mu$ g/ml chloramphenicol + 50  $\mu$ g/ml ampicillin) an equivalent amount of DNA yielded 1904 cfu, which represents a positive rate of 0.05%. Clones showed a variety of colony sizes  
25 ranging from 0.2-2 mm.

When greater concentrations of ampicillin were used, the fraction of positive clones decreased (see Table 2, B). In contrast, the plating efficiency of the control plasmid using CAM selection pCCHAsc1-ampR (A), which  
30 contains the wild Omp A signal, did not decrease appreciably at higher concentrations of ampicillin. The libraries formed under higher concentrations of ampicillin presumably contain clones with leader peptides capable of higher levels of secretion. In these experiments 10  $\mu$ g of  
35 pCCHAsc1-ampR was electroporated into 20  $\mu$ l

SUBSTITUTE SHEET (RULE 26)

-56-

electrocompetent *E. coli* DH10B and grown for one hour at 37°C in 1 mL S.O.C. (Sambrook, et al. supra). The cells were centrifuged briefly and resuspended in 400  $\mu$ L S.O.C. A  $10^4$  dilution was prepared in S.O.C. and 0.1 ml aliquots were plated on TB plates containing 34  $\mu$ g/ml chloramphenicol with the indicated concentration of ampicillin. This transformation strategy was compared to a transformation using 4  $\mu$ L (approximately 1  $\mu$ g) of signal peptide library DNA electroporated into 80  $\mu$ L electrocompetent *E. coli* DH10B and processed as described for the control pCCHAscl-ampR plasmid except that the  $10^4$  dilution step was eliminated.

Table 2

Restriction of the signal peptide library on increasing concentrations of ampicillin

	Clone/Library	[amp] (μg/mL)	cfu/plate	(+) cfu by 111In screen	% positive	
20	A	pCCHAsclampR	50	254	199	78.3
			100	218	149	68.3
			200	319	201	63.0
25			400	211	74	35.1
	B	SP Library	50	640	107	16.7
			100	313	98	31.3
			200	118	42	35.5
30			400	12	6	50.0

The library size was determined by electroporating 1  $\mu$ L of library DNA and plating 90% of the electroporated cells onto media containing 34  $\mu$ g/ml chloramphenicol and 100  $\mu$ g/ml ampicillin and by plating serial dilutions of the remaining 10% of the electroporated cells onto media containing only chloramphenicol (34

SUBSTITUTE SHEET (RULE 26)

-57-

µg/ml). Under the latter conditions all clones produced colonies. The quality of the resulting signal peptide library was verified by DNA sequencing of the mutated region of four nonselected clones. In all clones the mutations were correctly incorporated. The composition of the mixed bases was 31% A, 19% G, 25% C and 25% T where the sample number was 119.

Library clones that were able to grow on 50 µg/ml or greater ampicillin were further verified by colony lift and by hapten binding assays. One possible complication of ampicillin selection can be deletion of all or some of the scCHA insert. Another possibility is that the EIPCR reaction could incorporate mutations in the antibody sequence and thereby eliminate antibody activity. In fact some errors were identified when individual clones were later sequenced. To examine these possible complications, two approaches were used. First, digestion of DNA from positive clones (grown with 200 µg/ml ampicillin) with Xba I and Pst I indicated that about 70% of the clones contained the full length 800 scCHA insert.

Second, a large number of colonies were grown under selective conditions, transferred to nitrocellulose and probed with the radioactive hapten <sup>111</sup>In-EOTUBE. Labeled chelate was prepared and colony lifts performed as described by Stemmer, et al. (*BioTechniques* 14:256-265, 1993). Briefly, colony lifts of 23cmx23cm plates with 0.3-1 x 10<sup>5</sup> colonies were prepared using BA83 nitrocellulose filters (Schleicher and Schuell, Keene, New Hampshire). The filters were blocked by incubation in 3% non-fat milk in 25 mM Tris-HCl pH 7.5 for 10 minutes, washed with 25 mM Tris, followed by incubation in 25 mM Tris containing 50 °Ci of chelated <sup>111</sup>Indium per filter for 1 hour at room temperature. The filters were then washed with 25 mM Tris for a total of 15 minutes, dried and exposed to Kodak X-Omat AR autoradiography film for several hours.

SUBSTITUTE SHEET (RULE 26)

-58-

Approximately 30-50% of the colonies yielded a positive signal in this assay, indicating that a significant number of clones were not expressing active scCHA antibody (see Table 2, supra). Loss of scCHA activity also occurred in the control plasmid, though at a lower frequency. Clones yielded positive scCHA signals of varying intensities. Twelve clones yielding a high intensity signal were selected for further work and were purified by a secondary and tertiary colony screen. These positive clones were removed from the primary screen plates and diluted into fresh media, replated and reassayed. A single positive colony from this secondary screen was replated and reassayed in a tertiary screen, which yielded clonal material, since 100% of the colonies yielded a positive signal. The results of representative colony lifts shown demonstrated that the scCHA signal was stable during successive rounds of growth (data not shown).

In order to study protein expression in the cells, the cells were grown at 30°C in Terrific Broth (see supra) containing 34 µg/ml chloramphenicol and 50 µg/ml ampicillin in baffled shaker flasks at 250 rpm for 24 hours. Following cell growth, the periplasmic fraction of *E. coli* was prepared and isolated using the methods described by Witholt, et al. (*Anal. Biochem.* 74: 160-170, 1976, hereby incorporated by reference).

Determination of antibody expression levels was performed as previously described (Stemmer, et al., *BioTechniques*, 1993 supra). The quantitation of the active fusion protein was based on a comparison to the CHA 255 standard. Isolated clones were further characterized to demonstrate the production of an active fusion protein. To quantitate the amount of fusion protein recognizing the labelled hapten, 50 µl of periplasmic extract was incubated with 50 µl of carrier free <sup>111</sup>In-EOTUBE for 30 minutes. The mixture was then centrifuged through an ultrafree column

SUBSTITUTE SHEET (RULE 26)

-59-

(10,000 MW cutoff) and the eluate was counted. The % bound was expressed as the (input cpm-eluate cpm)/input cpm. The number of sites (in pmol) was determined from a standard curve utilizing CHA255 antibody. The data of Table 3 indicates that periplasmic supernatants from several clones bound to the radioactive hapten <sup>125</sup>IIn.

Table 3

10 Hapten binding and  $\beta$ -lactamase activity of protein containing mutant signal peptides

	clone	pmol site/ml	ng/ml	$\Delta$ AU/min/ml	ng protein/ml
15	CHA	12	960.0	NA	NA
	$\beta$ -lac STD	NA	NA	3.038	680
20	ompA				
	Fusion protein	1.13	59.9	1.075	481
	E3	0.009	0.45	0.104	45.6
25	E5	0.076	4.04	0.079	34.66
	E6	0.638	33.8	1.409	618.1
	E9	0.098	5.17	0.089	39.05
	E10	0.009	0.5	0.371	162.76
	E12	0.131	6.93	0.697	305.8
30	E13	0.33	17.47	0.184	80.72
	E14	0.36	19.09	0.086	37.7
	E15	0.079	4.19	0.074	32.46
	E16	0.007	0.4	0.225	98.7
	E17	0.012	0.65	0.087	38.16
35	E18	0.043	2.29	0.083	36.41

SUBSTITUTE SHEET (RULE 26)

-60-

E19	0.032	1.7	0.081	35.53
E20	0.014	0.76	0.074	32.5
E21	0.03	1.58	0.104	45.6
E22	0.035	1.85	0.100	43.9

5

Since library clones were isolated from ampicillin containing plates, one would predict that periplasmic supernatants should also contain significant  $\beta$ -lactamase enzymatic activity. The activity of the  $\beta$ -lactamase portion of the fusion protein was determined by the quantitative PADAC (2-[(N,N-dimethylanilin-4-yl)azo]pyridinium 3'-cephalosporin) assay (Meyer, et al., *Bioconjugate Chem.* 3:42-48, 1992, which is hereby incorporated by reference). 100  $\mu$ l aliquots of periplasmic supernatant were incubated at 37°C with  $A_{570} = 0.5$  of PADAC in a 2 ml stirred cuvette and the decrease in the  $A_{570}$  was monitored over time.  $\beta$ -lactamase activities were expressed as a change in absorbance unit/min/ml and mass determinations were made based on a commercially available *E. coli*  $\beta$ -lactamase (Sigma St. Louis, MO) standard of known concentration assayed under identical conditions. Results are provided in Table 3. The data reveals that all of the clones tested had significant  $\beta$ -lactamase activity. Interestingly, mutant E6 had the highest activity, above that of the control fusion protein. The use of a  $\beta$ -lactamase standard permitted calculation of the approximate mass of fusion protein produced.

Based on the data in Table 3, there seems to be no clear direct correlation of antibody binding activity with  $\beta$ -lactamase activity. For example, some clones like E13 and E14 showed moderate levels of antibody activity but very low levels of beta lactamase activity. Also, clones like E10, E12 and E16, showed moderate levels of beta lactamase activity while exhibiting very little antibody activity. Only clone E6 showed levels of antibody and

SUBSTITUTE SHEET (RULE 26)

-61-

enzyme activity that were close to that for the control Omp A clone. Because there is the possibility of obtaining mutations in any PCR technology that may affect the activity of antibody or enzymes, we took several of the

5 periplasmic preparations and evaluated them for their expression levels by Western blotting (Sambrook; et al. supra) since this method should be less sensitive to mutation (see Stemmer, et al. *BioTechniques*, 1993 supra). Samples were developed with purified rabbit polyclonal

10 antibodies to CHAsFv. The antibodies were obtained from rabbit polyclonal antiserum. Rabbits were immunized by a traditional procedure (*J. Immunol.* 116:1306-1312, 1975) that employed multiple intramuscular injections of CHA255 monoclonal antibody (100  $\mu$ g amounts) in Freund's adjuvant.

15 Serum obtained from the rabbits were processed to isolate a purified polyclonal antibody preparation to CHA255 Fv by absorption and elution from an immunoadsorbant column that was conjugated with CHA255 chimeric antibody. The CHA255 chimeric antibody was cloned and produced as described

20 previously (*J. Immunol.* 145:1200-1204, 1990 and *Protein Expr. and Purif.* 2:75-82, 1990). CHA255 chimeric antibody was conjugated to CNBr preactivated Sepharose 4B according to the manufacturers instructions (BioRad Labs, Richmond, CA). The antibody in the rabbit polyclonal antiserum were

25 applied to the column and eluted as described previously (*J. Immunol.* 116:1306-1312, 1975). This purified polyclonal antibody preparation was used in Western blots in 100  $\mu$ g/ml concentrations. The results of the Western Blot experiments indicated that E5 was now the best

30 producer of the mutants tested from Table 3. In this experiment, the Omp A control clone produced higher levels of protein than the mutants. Expression of E5 was higher than that of E6, with E12 and E10 slightly lower in expression than E6 (data not shown).

SUBSTITUTE SHEET (RULE 26)

-62-

In addition, several of the selected library E clones were subjected to DNA sequencing to determine the sequence of the DNA that coded for the signal peptide (see Table 4). A number of the clones contained stop codons within their sequence. In particular, clone E6, which directed increased protein expression in the antibody and enzyme activity tests and was second only to E5 in the Western blot, contained several stop codons in the signal sequence.

Table 4

AMINO ACID SEQUENCES OF SIGNAL PEPTIDE LIBRARY E

15	LEADER	SEQUENCE		
	OmpA (SEQ ID NO: 11)	MKKTAIAIAV	ALAGFATVAQ	A
	E3 (SEQ ID NO: 12)	MLYPGEMLCM	QNFI TLII EI	YI
	E5 (SEQ ID NO: 13)	MGM YKLRL LI	DCLMLYVCLT	DT
	E6 (SEQ ID NO: 14)	MNILL EN*M*	RTHDMLIDGT	I
20	E9 (SEQ ID NO: 15)	MGM YKLRL L	DCLMLYVCST	DT
	E10 (SEQ ID NO: 16)	MTSVCKITVL	ILFXTGT YNI	LG
	E12 (SEQ ID NO: 17)	MRNTYV'YXKY	GRTEWLVEQI	NL
	E13 (SEQ ID NO: 18)	MvQCERXIYH	YvQRIREYFX	FE
	E14 (SEQ ID NO: 19)	MvGICTGIGV	GILL LLLLVG	QV
25	E15 (SEQ ID NO: 20)	MLGAMEFKKL	SYAILVLSHT	PT
	E16 (SEQ ID NO: 21)	MDQGIKYRLE	KWRGRISLIR	VF
	E17 (SEQ ID NO: 22)	MQRRGIRWAY	VIYLSMFFGA	C
	E18 (SEQ ID NO: 23)	MLGVLTMPCD	TMVEDLACDY	MS

\* = stop codon

30 X = unknown amino acid

Because the EIPCR method involves amplification of the entire plasmid, there is always the possibility that one can have mutational errors occurring in any location in the vector of the selected library clones. Thus, it is

SUBSTITUTE SHEET (RULE 26)

-63-

always recommended that after the DNA sequence is determined for the mutated library segment, the selected new sequence is recloned using new DNA oligonucleotides with standard cloning procedures to confirm that the sequence is in fact directing the expected effect identified from the originally selected mutant sequence. In the case of the signal sequence library E, we chose to reclone E5 and E10 signal sequences in place of the natural  $\beta$ -lactamase signal sequence and compare expression with Omp A driven  $\beta$ -lactamase. To accomplish this, oligonucleotides containing the new sequences were cloned into the vector pCLA3ampR (Figure 3). This PUC derived vector contains a constitutively expressed chloramphenicol gene, LacI, and the  $\beta$ -lactamase gene. Expression of  $\beta$ -lactamase is driven by the Lac promoter and is controlled by the Omp A RBS sequence and Omp A signal sequence. The activity of the Lac promoter is negatively regulated by the LacI gene product, the Lac repressor.

Construction of pCLA3ampR E5 was accomplished by PCR using primers B947 (SEQ ID NO: 24) and 940 (5' and 3' primers respectively, see Figure 9) with the pCLA3ampR template using standard PCR conditions as disclosed supra. The product was cut with Xba I and Pst I. The PCR product was then ligated back into pCLA3ampR, which had previously been digested with Xba I and Pst I. The construction of pCLA3ampR E10 was accomplished in the same way, except that the 5' primer was B946 (SEQ ID NO: 25).

In order to determine the expression level of  $\beta$ -lactamase, clones were grown in Terrific Broth containing 30  $\mu$ g/ml chloramphenicol in baffled shaker flasks rotating at 250 rpm for 24 hours at 30°C. IPTG (Boehringer Mannheim, Indianapolis, IN) was added (final culture concentration 0.4 mM) to the cultures when the OD<sub>600</sub> reached 0.8 to induce the expression of the  $\beta$ -lactamase gene. The periplasmic fraction was obtained as described by Witholt,

SUBSTITUTE SHEET (RULE 26)

-64-

et al. (supra). The periplasmic samples were analyzed for  $\beta$ -lactamase activity using the PADAC assay described earlier. The results indicated that the E5 and E10 signal sequences were equivalent to Omp A in their ability to  
5 express  $\beta$ -lactamase from the pCLA3 vector (Table 5). It should be noted that the DNA sequence coding for the signal peptides of clones E5 and E10 depicted in Table 5, were determined. The sequence of E10 was as previously determined, but E5 showed a T to C base change at position  
10 8, resulting in a change of the third amino acid of E5 from a methionine to a threonine. The results in Table 5 indicate that this single amino acid change in E5 did not adversely affect its ability to function as a signal  
15 sequence.

15

-65-

Table 5

	Clone	Exp. 1	Exp. 2	Exp. 3	Average/ (stdev)	% of Control
			AU/min /ml x 10 <sup>-3</sup>			
5	pCLA3 ampR Control	19	23	22	21.3/ (2.1)	100
	RBS 1 short			40		188
10	RBS 2 short	46	48	35	43/(7.0)	202
	RBS 2 long	40	47	38	41.7/ (4.7)	196
15	RBS 8 short	57	71	59	62.3/ (7.6)	293
	RBS 12	39	47	36	41.7/ (3.5)	196
	Signal peptide E 5		20	22	21/(1.0)	99
20	Signal peptide E 10		21	23	22/(1.0)	103
	Blank		0	0	0	0

25                   The library generated E5 and E10 signal sequences  
were then compared for structural properties with selected

**SUBSTITUTE SHEET (RULE 26)**

-66-

signal sequences derived from *E. coli* proteins (Omp A, DGAL, LamB, ELT6, Omp T, Pel B, PhoA). These signal sequences, as well as a variety of others are available from the GenBank database. The program Protean (DNASTar, Madison, WI) was used to evaluate alpha helix vs beta sheet regions (Garnier-Robson and Chou-Fasman methods) as well as the hydrophilicity index (Kyte-Doolittle method). Both E5 and E10 displayed the characteristic hydrophobic core of amino acids seen in all of the *E. coli* derived signal sequences (Figure 4). Interestingly, the E10 sequence contained no alpha helix regions. Alpha helix content varied with all of the native *E. coli* signal sequences evaluated. Also, an interesting feature of the E5 signal sequence was that the final 3 residues of the sequence, Thr, Asp, Thr, correspond to those designated in the art as a -3,-1 box which provides the recognition sequence for cleavage by the *E. coli* signal peptidase (Protein Targeting, supra). The sequence in the -3,-1 box for E5 differs from all known *E. coli* sequences. The presence of this unique -3,-1 sequence is not detrimental to the ability of E5 to export a functional  $\beta$ -lactamase protein to the periplasm.

To demonstrate that the fusion proteins expressed with mutant signal peptides were processed to the correct size, periplasmic extracts were fractionated on SDS-polyacrylamide gels. The proteins were transferred to nitrocellulose and probed with an anti-CHA Fv antibody. Results from the Western blots indicated that the fusion proteins isolated from the bacterial clones were 53 kD. This is close to the predicted size of the fusion protein.

#### Example 2

##### Production of Novel Signal Sequences of Variable Length

Figure 5 illustrates the use of primers B755, B756, B757, and B758 (corresponding to SEQ ID NOS: 26 - 30

-67-

respectively) that were used in EIPCR reactions to produce signal peptide libraries of varying length. These libraries were constructed using primer B755 and one of the following primers: B756, corresponding to the random 12 aa signal peptide library; B757, corresponding to the random 16 aa signal peptide library; and B758, corresponding to the random 20 aa peptide library. In this example, the varying length libraries were constructed such that the random signal sequences were incorporated directly onto the amino terminus of the gene sequence encoding  $\beta$ -lactamase (amp-r). The resulting plasmids are diagrammed in Figure 5 as pCampR SP-12, pCampR SP-16 and pCampR SP-20.

Following EIPCR, and using the methods outlined in Example 1, several clones from each library were prepared and the DNA was analyzed by restriction endonuclease digestion to verify the predicted size of the constructs. In the table below, active clones are defined as those clones which grow on 34  $\mu$ g/ml chloramphenicol plus 10  $\mu$ g/ml ampicillin. The following percentages of active clones were identified. Active clones are defined as those which grown on 34  $\mu$ g/ml chloramphenicol plus 10  $\mu$ g/ml ampicillin):

	<u>Library</u>	<u>% Active Colonies</u>
25	12 aa	0.0028%
	16 aa	0.0057%
	20 aa	0.0141%

The clones were also analyzed by PADAC assay (outlined in Example 1) and clones with high  $\beta$ -lactamase activity were selected for sequencing. The results indicated that as the length of the signal sequence increases, the percentage of active clones within the library also increases. Moreover, while it is possible to

-68-

obtain shorter signal sequences that were active, signal sequences of at least 20 aa have an increased likelihood of being functional.

### 5 Example 3

#### Creation of Novel Ribosome Binding Sites in *E. coli*

To apply this invention to a second regulatory region, we deleted the DNA sequence in an expression vector coding for a ribosome binding site (RBS), and generated a  
10 random DNA library in its place. Since the RBS sequence initiated translation of  $\beta$ -lactamase, library clones which contained a useful RBS sequence were selected by growing the library on ampicillin. The RBS library was prepared in the pCampR vector (Figure 6). This is a pUC derived vector  
15 which has a constitutively expressed chloramphenicol resistance gene and expressed  $\beta$ -lactamase under the control of the Lac promoter. In addition, translation of the  $\beta$ -lactamase gene was accomplished using a ribosome binding site (GAGG) from the Omp A gene (Omp A sequence from -18 to  
20 -1) followed by the Omp A signal sequence (+1 to +63) to direct the  $\beta$ -lactamase protein to the periplasm. The pCampR vector has a wild type DNA  $\beta$ -lactamase sequence. In order to use the Bsa I enzyme for the EIPCR library reaction, a Bsa I site in the  $\beta$ -lactamase gene was first  
25 removed. Removal was accomplished by standard EIPCR methodology (disclosed above) using the oligonucleotide primers 939 (SEQ ID NO: 3) and 940 (SEQ ID NO: 4) (see Figure 6). The resulting vector was designated pCampR\_BSA.

Next, the entire region containing the Omp A  
30 derived ribosome binding site was removed from pCampR\_BSAI. Sixteen bases were removed from position -16 to -1 by a single round of EIPCR using standard conditions and primers B667 (SEQ ID NO: 30) and B668 (SEQ ID NO: 31) (See Figure 6). The resulting vector was designated pCampR\_RBS and was  
35 unable to grow on plates containing 100  $\mu$ g/ml ampicillin.

-69-

To generate the RBS library, pCampRARBS was used as the template in an EIPCR reaction containing primers B765 (SEQ ID NO: 32) and B766 (SEQ ID NO: 33) (See Figure 6). Primer B766 contained a sequence of 16 NTPs (25% mixture of each base) so as to exactly replace the prior deleted 16 base sequence with a random DNA sequence. The EIPCR library reaction was accomplished using the EIPCR conditions as disclosed by Stemmer, et al. (*BioTechniques* 13:114-220, 1992). This reaction included low amounts of template (0.5-1 ng/100  $\mu$ l reaction mixture). This is in contrast to Example 1, which required much larger amounts of DNA template (up to 50 ng/100  $\mu$ l reaction mixture). The amount of template DNA that needs to be used in a particular EIPCR library reaction can be experimentally determined by those with skill in the art. As illustrated in Example 1, deletion of the entire functional region of the DNA prior to introducing the library sequence eliminates the concern for wildtype clones contaminating the library and enables the use of large amounts of template DNA for EIPCR, if indeed large amounts are required.

The EIPCR reaction material was processed to fill in the ends with Klenow and T4 DNA polymerase, cut with Bsa I, and ligated as described in Example 1 except that after the Bsa I digest, the DNA was electrophoresed on an agarose gel and the band containing the DNA was cut out and processed using GeneClean (Bio101, San Diego, CA), according to the manufacturer's instructions, to separate the digested from undigested fragments.

The ligated RBS library DNA was electroporated by adding 1  $\mu$ l (50 ng) of DNA to 20  $\mu$ l of electrocompetent DH10-B MAX cells (BRL, Bethesda, MD) according to the manufacturers instructions. Under conditions to select for the plasmid marker (34  $\mu$ g/ml chloramphenicol), the library size was estimated to be  $1.7 \times 10^6$  colony forming units/ $\mu$ g ligated DNA. Under dual selective conditions (34  $\mu$ g/ml

SUBSTITUTE SHEET (RULE 26)

-70-

chloramphenicol + 50  $\mu$ g/ml ampicillin), an equivalent amount of DNA was estimated to produce 24,040 colony forming units (cfu), which indicates that 1.4% of the colonies contained potential mutants. When greater

5 concentrations of ampicillin were used, the fraction of positive colonies on the double selected plates decreased precipitously (Table 6). In contrast, the plating efficiency of the control plasmid, pCampR $\Delta$ BSA, did not decrease appreciably until 400 or 800 ng/ml ampicillin was

10 used. Referring to Table 6, sixteen library colonies from the dual selective plates (800  $\mu$ g/ml ampicillin + 34  $\mu$ g/ml chloramphenicol) were picked and inoculated into LB media and grown overnight. In the morning, cells were harvested and periplasmic fractions were produced according to the

15 above referenced methods. PADAC as says to measure  $\beta$ -lactamase activity showed a wide range of activity (Table 7), with several clones exhibiting activities higher than what was obtained with the wildtype Omp A RBS in the pCampR $\Delta$ BSA vector. After several repeat cultures and PADAC

20 assays, it was clear that RBS clones 1, 8 and 12 showed higher levels of expressed  $\beta$ -lactamase activity than the wildtype Omp A RBS clones pCampR $\Delta$ BSAI (see Table 7).

-71-

Table 6

## ELECTROPORATION COLONIES

5	[AMP] ( $\mu\text{g/ml}$ )	RBS Library	Control pCamp $\Delta\text{BSA}$
	50	5000	1894
10	100	760	2007
	200	81	1564
	400	41	914
	800	22	211
15			

Table 7BETA LACTAMASE ACTIVITIES (PADAC) FOR RIBOSOME BINDING SITE  
LIBRARY CLONES

20	Selected *		Not Selected **	
	Clone#	Activity*** (in $\Delta\text{AU/min/ml}$ )	Clone#	Activity* ( $\Delta\text{AU/min/ml}$ )
25	RBS 1	5.58♦	RBS 11	1.95
	RBS 2	1.65♦	RBS 12	5.18
	RBS 3	2.72	RBS 13	0.012
	RBS 4	3.51♦	RBS 14	0.021
30	RBS 5	3.78	RBS 15	0.016
	RBS 6	2.09♦	RBS 16	0.005
	RBS 7	0.95		
	RBS 8	12.76♦		
	RBS 9	2.35♦		
35	RBS 10	3.54♦		

SUBSTITUTE SHEET (RULE 26)

-72-

- \* clones derived from plates containing 800  $\mu\text{g/ml}$  AMP,  
34  $\mu\text{g/ml}$  CAM
- \*\* clones derived from plates containing 34  $\mu\text{g/ml}$  CAM
- 5 \*\*\* PADAC activity for the wildtype ompA RBS sequence  
(pCampRBSAI) is 3.31  $\text{AU/min/ml}$
- ◆ These clones had an insertion of 17 bases added 3' of  
the library RBS sequences as measured by an enlarged  
Xba I, Eco RI restriction fragment

10

Purified DNA from the selected RBS clones were initially evaluated for key restriction sites. It was noticed that some of the library clones (RBS 1, 2, 4, 8-10) had acquired an unexpected mutation during formation of the library. Although it is not unusual to pick up errors during PCR techniques, it is important to note that the errors incorporated into the constructs disclosed below had no impact on the activities of the RBS sequences that were isolated from the library. This is known with certainty since we later recloned the selected sequences without the errors and found that the improved RBS sequences were as active as they were in the initial clones from the library.

The error we found resulted in an insertion of about 17 nucleotides arising from incomplete Bsa I digestion of one of the ends of the PCR library reaction, causing a blunt ended ligation. The insertional error did not appear to be related to selection conditions of the library, since it was seen in both the amp/CAM selected (50%, N=24) and CAM only selected clones (19%, n=16). A comparison of  $\beta$ -lactamase activities of the ampicillin selected clones also showed no obvious correlations between the insertion error and the level of beta lactamase activity produced (Table 7).

In order to determine if the insertional mutation played a role in the RBS clones exhibiting increased levels

-73-

of expressed  $\beta$ -lactamase, the DNA coding sequence for the RBS regions in four clones were determined by DNA sequencing and new oligonucleotides containing these sequences were synthesized to enable expression analysis in a new vector that had not undergone EIPCR. The DNA sequence for RBS clones #1, 2, 8 and 12 are provided in Figure 8. As noted above, for clones 1, 2 and 8 there was an additional insertion due to improper cutting of the terminus of the EIPCR primer (inserted nucleotides shown in bold). In clones 1, 2, and 8, the full length RBS sequence containing the insertion is herein referred to as RBS long, while the RBS sequence without the insertion is called RBS short. DNA oligonucleotide primers coding for resistance, RBS 2 long and short, RBS 8 short and RBS 12 were synthesized. Their sequences are shown in Figure 9.

The vector used to retest the activity of these various new RBS sequences was pCLAampR (Figure 3), disclosed above. The various clones were produced by performing a standard round of PCR in which the various primers were included and pCLAampR was used as the template (Figure 8). The PCR products were filled in with Klenow and T4 DNA polymerase, followed by restriction digestion using Xba I and Pst I. The products were cloned into the pCLAampR vector that had been previously cut with the same restriction enzymes. The DNA was ligated using the methods provided in Example 1, and DH10B bacteria were transformed using the electroporation procedures disclosed above. Clones were selected for growth on chloramphenicol and grown overnight in Terrific Broth media + 34  $\mu$ g/ml chloramphenicol. The next day, cultures were seeded at  $OD_{600}=0.1$  in LB media + 34  $\mu$ g/ml chloramphenicol and grown at 30°C to  $OD_{600}=0.8$ . IPTG was then added to a final concentration of 0.4mM. The cultures were grown for an additional 24 hours at which time the cells were harvested and periplasmic fractions were produced as described in

SUBSTITUTE SHEET (RULE 26)

-74-

Example 1. The periplasmic samples were next subjected to PADAC assay for  $\beta$ -lactamase activity determination. The results from three separate sets of cultures are provided in Table 5 and indicate that in the Lac regulated vector  
5 all the RBS library sequences shown in Table 5 were more active than the wild type Omp A RBS sequence. Also, the same level of  $\beta$ -lactamase expression was obtained for both the long and short versions of RBS clone 2, which indicates that the additional insertion from the original EIPCR  
10 reaction played no role in initiating translation.

It should be noted that DNA sequencing was used to verify the recloned RBS sites in all the RBS clones shown in Table 5. All sequences were unchanged except for clone 12, in which a thymidine nucleotide was deleted from  
15 the sequence. It is known from the literature that a critical element within an RBS sequence is an area of three or more bases in the mRNA that hybridize with the 3' end of 16S rRNA in bacteria (also known as the Shine and Dalgarno sequence). The underlined bases in Table 8 show potential  
20 areas of mRNA complementarity to the 3' end of 16S rRNA. Potential Shine and Dalgarno sequences were identified for RBS clones 2 and 12, but not for clones 1 and 6.

It should be pointed out that the mutagenesis process described here is very powerful due to its ability  
25 to optimize sequences at both the RNA and the protein level. In the case of the RBS sequences, some clones were isolated with improved expression levels, even though the area of the RBS sequence contained a 5' unexpected insertion of 17 nucleotides derived from an uncut primer  
30 sequence. In fact, these selected RBS sequences proved themselves to function well, whether or not the accidental insertion was present downstream. In addition, some of these sequences are working efficiently even though they do not contain a potential Shine and Dalgarno type sequence.  
35 Thus, this data illustrates the ability of the library

SUBSTITUTE SHEET (RULE 26)

-75-

approach to randomly incorporate sequences, both in predicted locations and in unpredicted locations and to assess the effectiveness of the mutations on protein expression irrespective of the location of the mutations.

5

-76-

Table 8

## DNA Sequences from Original Ribosome Binding Site Library

## Clones

5    Sequence Names                    DNA Sequence (5'-3' ton strand)  
      Omp A RBS                        TAACGAGGCGCAAAAA  
      (SEQ ID NO: 34)

## Library RBS

10        #1 long    AGACACGTACAAACCAATGAAAGAGACCTATTTACTC  
      (SEQ ID NO: 35)

     #2 long                    AAGCAAAGTCCCGCGAATGAAAGAGACCTATTT  
      (SEQ ID NO: 36)

15        #8 long                    ACGTTTAAACAGACACATGAAAGAGACCTATTT  
      (SEQ ID NO: 37)

     #12                        GGAACTCAAAGGCACC  
      (SEQ ID NO: 38)

-----  
      \*                    DNA sequence is from the XBA1 site (-16 position  
                          relative to the ompA signal peptide sequence) to  
                          the start of the Omp A signal sequence (-1  
 25                    position).

     Bolder nucleotides indicate the DNA sequence  
      corresponding to an insertional error due to  
      blunt ended ligation of a EIPCR DNA primer that  
 30                    failed to be cut by BSA1.

     Underlined nucleotides indicate areas of  
      potential complementarity with the 3' end of *E.*  
      *coli* 16S ribosomal RNA. (ie. potential Shine  
 35                    Dalgarno sequences).

SUBSTITUTE SHEET (RULE 26)

-77-

## Example 4

## Creation of Novel Regulatory Sequences in Mammalian Cells

As an example of the application of this invention to the development of novel regulatory sequences in mammalian cells, this example provides a method for producing signal peptide libraries in mammalian cells for improved protein expression. Figure 7a illustrates the vector contemplated for use in this invention, pGCEMK. The vector uses the immunoglobulin Kappa Light Chain promoter (from the kappa light chain gene of the murine monoclonal antibody CEM 231) to direct transcription of a desired gene. In this example it is driving the native CEM kappa light chain variable region, but one skilled in the art will be aware that any natural or recombinant variable region could replace the native Variable region and the same process of signal sequence mutagenesis could be accomplished. Further, the entire light chain coding sequences could be replaced by those from any eukaryotic gene and a similar process carried out. In the present example, the aim of the mutagenesis is to determine if an alternate signal sequence can be obtained of equivalent or improved efficacy, as measured by expression of mature light chain.

Other components of the vector, pGCEMK, include the major intron from the human kappa light chain gene, containing its native enhancer sequence; the human kappa constant region, containing its native polyadenylation signal; a Xanthine-guanine phosphoribosyl transferase gene (gptR) under the control of an SV 40 promoter to permit selection of stable transformed cells; the bacterial colE1 origin of replication, to permit high copy number plasmid replication in *E. coli*. The plasmid, pGCEMK, was constructed from the plasmids pHF-1 and pMLCE-10, which have been deposited with the American Type Culture

-78-

Collection as Accession numbers 67,637 and 67,639,  
respectively. These plasmids and the construction pGCEMK  
are described in U.S. patent application Serial No.  
07/727,719, filed July 2, 1991, which is incorporated  
5 herein by reference in its entirety.

In this example the native signal sequence is  
removed. In one vector, this region is replaced with the  
first two codons of the native signal to allow efficient  
initiation of translation to take place. These are fused  
10 directly to the first codon of the Variable region. Random  
60-65 mer oligonucleotides are inserted into the vector by  
PCR. One skilled in the art has the option of conserving  
the signal intron in the resultant construct or deleting  
it, as long as the control containing the native signal has  
15 a comparable format. Following incorporation of the random  
nucleotides into the eukaryotic vector, the vector is  
introduced into suitable eukaryotic cells that are capable  
of expressing protein from the expression vector (for  
example, SP 2/0 cells, American Type Culture Collection,  
20 Rockville, MD). In this example, electroporation is used  
to mediate transfection; however, those skilled in the art  
will be readily able to select other equally appropriate  
methods for introducing their selected expression vector  
into a suitable cell type. Positive cells are preferably  
25 selected by growing the cells on media containing  
hypoxanthine, mycophenolic acid and xanthine (HMAX).

At this point, one skilled in the art could  
enrich for cells actually expressing protein at a  
detectable level, for instance, by Fluorescence Activated  
30 Flow Cytometry with Cell Sorting. In this example, cells  
expressing a detectable level of protein are identified by  
ELISA. Positive colonies are further quantitated for  
protein expression by a quantitative ELISA technique as  
compared with cells expressing protein under the control of  
35 the native regulatory sequence, and using the vector

SUBSTITUTE SHEET (RULE 26)

-79-

without random oligonucleotides added as a negative control. Cells expressing equal or greater levels of protein as compared with cells containing the native regulatory sequence are subjected to DNA sequencing, and  
5 identified as optimal signal sequences for the particular protein expressed.

#### Example 5

#### 10 Creation of Novel Protein Trafficking Signals in Mammalian Cells

A critical component of immunoglobulin (Ig) protein trafficking is the binding of the chaperon protein BiP to immunoglobulin heavy chain and the subsequent  
15 displacement of BiP by light chain (*Nature* 306:387-389, 1983). Secretion of the intact immunoglobulin molecule is not possible unless BiP is disassociated from the heavy chain. The kinetics of these interactions impact directly upon the level of expression of intact antibody. For  
20 example, if binding of heavy chain monomers by BiP is inefficient, non-bound heavy chain monomers can be secreted. These monomers compete with intact molecules and inhibit the secretion process. Alternatively, the inefficient displacement of BiP, by light chain, from BiP-  
25 associated heavy chain would slow the process of intact immunoglobulin secretion. The BiP binding regions within the immunoglobulin heavy chain have been narrowed to specific regions on the molecule (in  $C_{H1}$  and  $V_H$ ; *Blood* 79:2181-2195, 1992), encoded by specific regions of DNA in  
30 the gene. The methods of this invention can be used to identify optimized sequences that increase overall expression of intact Ig without increasing the level of light or heavy chain monomers (or homodimers).

A mammalian expression vector is selected that  
35 contains a heavy chain gene. Figure 8 illustrates one

-80-

heavy chain chimeric vector, pNCEMG1, that is useful for practicing the trafficking methods of this invention. This vector uses the Immunoglobulin heavy chain promoter to transcribe the desired gene inserted into a cloning site.

5 The vector includes an immunoglobulin heavy chain polyadenylation signal and includes a neomycin resistance gene. The neomycin resistance gene is under the control of the SV40 promoter, thereby permitting the selection of stable transformed cells in media containing the antibiotic

10 G418. The bacterial colE1 origin of replication permits high copy number plasmid replication in *E. coli*. Construction of this plasmid from plasmids pHG12 and PMHCE30, which have been deposited in the American Type Culture Collection (Accession Nos. 67,638 and 67,640,

15 respectively) as described in pending U.S. patent application Serial No. 07/727,319. In this example, the native CEM immunoglobulin heavy chain V-region sequence is incorporated into the vector, but one skilled in the art will be aware that any natural or recombinant variable

20 region could replace the native variable region and the same process of BiP binding region optimization could be carried out.

The normal BiP binding region located within the human gamma 1 constant region gene is operably linked to

25 the recombinant variable region in this vector. A second vector is now constructed that is identical to the first except that the BiP binding region is deleted from the C<sub>H1</sub> domain. Random 12-20 mer sequences are inserted into the vector using the PCR methods described in Example 1.

30 Following incorporation of the random oligonucleotides into the eukaryotic vector, the vector is introduced into suitable eukaryotic cells, which produce a light chain complementary to the heavy chain used (e.g. the cell lines produced in Example 4). The vector can be introduced into

35 cells using methods well known in the art. Here,

-81-

electroporation is used to mediate transfection. Positive cells are preferably selected by growing the cells in media containing G418.

At this point, one skilled in the art could  
5 enrich for cells actually expressing protein at a detectable level, for instance, by Fluorescence Activated Flow Cytometry with Cell Sorting. In this example, cells expressing a detectable level of protein are identified by ELISA. Positive cells are further quantitated for protein  
10 expression by a quantitative ELISA technique as compared with cells expressing the protein under the control of the native regulatory sequence, and using the vector without random oligonucleotides added as a negative control. Cells expressing equal or greater levels of intact antibody as  
15 compared with positive control cells and that do not express an increased level of free light or heavy chains are subjected to DNA sequencing to determine the DNA sequence of the trafficking signals most efficacious for expression of the immunoglobulin.

20 While particular embodiments of the invention have been described in detail, it will be apparent to those skilled in the art that these embodiments are exemplary rather than limiting, and the true scope of the invention is that defined in the following claims.

25

-82-

## SEQUENCE LISTING

(1) GENERAL INFORMATION:

5 (i) APPLICANT: Antelman, Douglas E.  
Wilson, Barry S.

10 (ii) TITLE OF INVENTION: METHOD FOR CREATING  
OPTIMIZED REGULATORY REGIONS AFFECTING  
PROTEIN EXPRESSION AND PROTEIN TRAFFICKING

(iii) NUMBER OF SEQUENCES: 43

15 (iv) CORRESPONDENCE ADDRESS:  
(A) ADDRESSEE: Eli Lilly and Company  
(B) STREET: Lilly Corporate Center  
(C) CITY: Indianapolis  
(D) STATE: Indiana  
20 (E) COUNTRY: United States of America  
(F) ZIP: 46285

(v) COMPUTER READABLE FORM:  
(A) MEDIUM TYPE: Floppy disk  
(B) COMPUTER: Apple MacIntosh  
25 (C) OPERATING SYSTEM: Apple System 7.0  
(D) SOFTWARE: Microsoft Word

(vi) CURRENT APPLICATION DATA:  
(A) APPLICATION NUMBER:  
30 (B) FILING DATE:  
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:  
(A) NAME: Gaylo, Paul J.  
35 (B) REGISTRATION NUMBER: 36,808

**SUBSTITUTE SHEET (RULE 26)**

-83-

(C) REFERENCE/DOCKET NUMBER: H-8589

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (317) 276-0756

5 (B) TELEFAX: (317) 276-3861

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

10 (A) LENGTH: 26 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

20

25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

ATATATTCTA GATAACGAGG CGCAA

26

(2) INFORMATION FOR SEQ ID NO: 2:

30

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 54 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

35 (D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

-84-

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

5 (iv) ANTI-SENSE: NO

10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

ATAAATTGGT CTCACCAGAA CCACCACCAC CACCTGCAGA GACAGTGACC  
AGAG

54

15 (2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 25 base pairs

(B) TYPE: nucleic acid

20 (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

25 (iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

30

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GGTCTCGCGG TATCATTGCA GCACT

35 25

-85-

## (2) INFORMATION FOR SEQ ID NO: 4:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 35 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

20 AATTGGTCTC GGAACCACGC TCACCGGCTC CAGAT

35

## (2) INFORMATION FOR SEQ ID NO: 5:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 53 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

SUBSTITUTE SHEET (RULE 26)

-86-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

5 ATAAATTGGT CTCATCTGGT GGTGGTGGTT CTCACCCAGA AACGCTGGTG  
AAA

53

(2) INFORMATION FOR SEQ ID NO: 6:

10 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 44 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

15

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

20 (iv) ANTI-SENSE: NO

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATATATGGAT CCGGTACCCT ATCATTACCA ATGCTTAATC AGTG

44

30 (2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 32 base pairs  
(B) TYPE: nucleic acid  
35 (C) STRANDEDNESS: single

SUBSTITUTE SHEET (RULE 26)

-87-

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

5 (iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

15 ATAATTAGGT CTCCCAGGCT GTGGTGACCC AG

32

(2) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:

20 (A) LENGTH: 32 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

30

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

35 ATAATTAGGT CTCACCTGTT TTTGCGCCTC GT

32

SUBSTITUTE SHEET (RULE 26)

-88-

## (2) INFORMATION FOR SEQ ID NO: 9:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 31 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- (iv) ANTI-SENSE: NO

15

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

20 ATAATTAGGT CTCAAGGCTG TGGTGACCCA G

31

## (2) INFORMATION FOR SEQ ID NO: 10:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 100 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- (iv) ANTI-SENSE: NO

35

SUBSTITUTE SHEET (RULE 26)

-89-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

5 AATAATAGGT CTCTGCCTGN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN  
NNNNNNNNNN 60

NNNNNNNNNN NNNNNNNNNN NNCATTTTTT GCGCCTCGTT 100

10 (2) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 21 amino acids  
(B) TYPE: amino acid  
15 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

20 (iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal  
25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:  
30

Met Lys Lys Thr Ala Ile Ala Ile Ala Val Ala Leu Ala Gly Phe  
Ala  
1 5 10 15

SUBSTITUTE SHEET (RULE 26)

-90-

Thr Val Ala Gln Ala

20

## (2) INFORMATION FOR SEQ ID NO: 12:

5

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

10

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

15

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

25 Met Leu Tyr Pro Gly Glu Met Leu Cys Met Gln Asn Phe Ile Thr  
Leu

1

5

10

15

Ile Ile Glu Ile Tyr Ile

30

20

## (2) INFORMATION FOR SEQ ID NO: 13:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids

35

SUBSTITUTE SHEET (RULE 26)

-91-

(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: peptide  
  
(iii) HYPOTHETICAL: NO  
  
(iv) ANTI-SENSE: NO  
10 (v) FRAGMENT TYPE: N-terminal

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

Met Gly Met Tyr Lys Leu Arg Leu Leu Ile Asp Cys Leu Met Leu  
Tyr

20 1 5 10 15

Val Cys Leu Thr Asp Thr  
20

25 (2) INFORMATION FOR SEQ ID NO: 14:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 21 amino acids  
(B) TYPE: amino acid  
30 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

35 (iii) HYPOTHETICAL: NO

SUBSTITUTE SHEET (RULE 26)

-92-

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

5

(ix) FEATURE:

(A) NAME/KEY: Stop Codon

(B) LOCATION: 8, 10

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

Met Asn Ile Leu Leu Glu Asn Xaa Met Xaa Arg Thr His Asp Met  
Leu

15    1                                    5                                    10                                    15

Ile Asp Gly Thr Ile

20

20 (2) INFORMATION FOR SEQ ID NO: 15:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 21 amino acids

(B) TYPE: amino acid

25 (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

30 (iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

35

SUBSTITUTE SHEET (RULE 26)

-93-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

5 Met Gly Met Tyr Lys Leu Arg Leu Leu Asp Cys Leu Met Leu Tyr  
Val  
1 5 10 15

Cys Ser Thr Asp Thr  
10 20

(2) INFORMATION FOR SEQ ID NO: 16:

(i) SEQUENCE CHARACTERISTICS:  
15 (A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

20 (ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO  
25

(v) FRAGMENT TYPE: N-terminal

30

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

Met Thr Ser Val Cys Lys Ile Thr Val Leu Ile Leu Phe Xaa Thr  
Gly  
35 1 5 10 15

SUBSTITUTE SHEET (RULE 26)

-94-

Thr Tyr Asn Ile Leu Gly

20

## (2) INFORMATION FOR SEQ ID NO: 17:

- 5 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

10

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

15

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

20

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

25 Met Arg Asn Thr Tyr Val Tyr Xaa Lys Tyr Gly Arg Thr Glu Tr  
Leu

1

5

10

15

Val Glu Gln Ile Asn Leu

20

30

## (2) INFORMATION FOR SEQ ID NO: 18:

- 35 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid

SUBSTITUTE SHEET (RULE 26)

-95-

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

5

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

10

(v) FRAGMENT TYPE: N-terminal

15

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18:

Met Val Gln Cys Glu Arg Xaa Ile Tyr His Tyr Val Gln Arg Ile  
Arg

1

5

10

15

20

Glu Tyr Phe Xaa Phe Glu

20

(2) INFORMATION FOR SEQ ID NO: 19:

25

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 22 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: single

30

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

35

SUBSTITUTE SHEET (RULE 26)

-96-

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 19:

Met Val Gly Ile Cys Thr Gly Ile Gly Val Gly Ile Leu Leu Leu  
Leu  
1 5 10 15  
10 Leu Val Val Gly Gln Val  
20

(2) INFORMATION FOR SEQ ID NO: 20:

15

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
20 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

25

(iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

30

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

SUBSTITUTE SHEET (RULE 26)

-97-

Met Leu Gly Ala Met Glu Phe Lys Lys Leu Ser Tyr Ala Ile Leu  
Val

1                      5                      10                      15

5 Leu Ser His Thr Pro Thr

20

(2) INFORMATION FOR SEQ ID NO: 21:

10 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

15

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

20 (iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:

30 Met Asp Gln Gly Ile Lys Tyr Arg Leu Glu Lys Trp Arg Gly Arg  
Ile

1                      5                      10                      15

Ser Leu Ile Arg Val Phe

20

35

SUBSTITUTE SHEET (RULE 26)

-98-

## (2) INFORMATION FOR SEQ ID NO: 22:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 21 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (iii) HYPOTHETICAL: NO
- (iv) ANTI-SENSE: NO
- (v) FRAGMENT TYPE: N-terminal

## 20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:

Met Gln Arg Arg Gly Ile Arg Trp Ala Tyr Val Ile Tyr Leu Ser  
Met

1 5 10 15

Phe Phe Gly Ala Cys  
20

## 30 (2) INFORMATION FOR SEQ ID NO: 23:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 22 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

-99-

(ii) MOLECULE TYPE: peptide

(iii) HYPOTHETICAL: NO

5 (iv) ANTI-SENSE: NO

(v) FRAGMENT TYPE: N-terminal

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:

15 Met Leu Gly Val Leu Thr Met Pro Cys Asp Thr Met Val Glu Asp  
Leu

1 5 10 15

Ala Cys Asp Tyr Met Ser

20

20

(2) INFORMATION FOR SEQ ID NO: 24:

(i) SEQUENCE CHARACTERISTICS:

25 (A) LENGTH: 120 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

30

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

35

SUBSTITUTE SHEET (RULE 26)

-100-

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:

TAATTATTCT AGAATGGGGA TGTACAAGCT GAGACTACTG ATAGATTGCC  
5 TAATGCTTTA 60  
TGTTTGTTCA ACCGATACCG GAATTCCGGG TCACCCAGAA ACGCTGGTGA  
AAGTAAAAGA 120

## (2) INFORMATION FOR SEQ ID NO: 25:

10

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 120 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
15 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

20

(iv) ANTI-SENSE: NO

25

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

TAATTATTCT AGAATGACTT CTGGGAGCAA AATAACTGTT CTTATACTTC  
TTTTGACTGG 60  
30 TACATACAAT ATATTAGGGG GAATTCCGGG TCACCCAGAA ACGTCGGTGA  
AAGTAAAAGA 120

## (2) INFORMATION FOR SEQ ID NO: 26:

35

- (i) SEQUENCE CHARACTERISTICS:

SUBSTITUTE SHEET (RULE 26)

-101-

- (A) LENGTH: 31 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

10

(iv) ANTI-SENSE: NO

15

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 26:

ATAATAGGTC TCCACCCAGA AACGCTGGTG A

31

(2) INFORMATION FOR SEQ ID NO: 27:

20

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 72 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

25

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

30

(iv) ANTI-SENSE: NO

35

SUBSTITUTE SHEET (RULE 26)

-102-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

ATAATAGGTC TCTGGGTGNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN  
NCATTTTTTTG 60  
5 CGCCTCGTTA TC 72

(2) INFORMATION FOR SEQ ID NO: 28:

(i) SEQUENCE CHARACTERISTICS:  
10 (A) LENGTH: 84 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO  
20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 28:

25 ATAATAGGTC TCTGGGTGNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN  
NNNNNNNNNN 60  
NNNCATTTTTT TGCGCCTCGT TATC 84

(2) INFORMATION FOR SEQ ID NO: 29:

30 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 96 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
35 (D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

-103-

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

5 (iv) ANTI-SENSE: NO

10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 29:

ATAATAGGTC TCTGGGTGNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN  
NNNNNNNNNN 60  
NNNNNNNNNN NNNNNCATT TTTGCGCCTC GTTATC 96

15 (2) INFORMATION FOR SEQ ID NO: 30:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 31 base pairs

20 (B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

25 (iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 30:

35 ATAAATAGGT CTCATGAAAA AGACAGCTAT C 31

SUBSTITUTE SHEET (RULE 26)

-104-

## (2) INFORMATION FOR SEQ ID NO: 31:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 31 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

5

- (ii) MOLECULE TYPE: cDNA

10

- (iii) HYPOTHETICAL: NO

- (iv) ANTI-SENSE: NO

15

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 31:

20 ATAAATAGGT CTCCTCTAGA AATTGTGAAA T

31

## (2) INFORMATION FOR SEQ ID NO: 32:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 32 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

25

- (ii) MOLECULE TYPE: cDNA

30

- (iii) HYPOTHETICAL: NO

- (iv) ANTI-SENSE: NO

35

SUBSTITUTE SHEET (RULE 26)

-105-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 32:

5 ATAAATAGGT CTCATGAAAA AGACAGCTAT CG 32

(2) INFORMATION FOR SEQ ID NO: 33:

(i) SEQUENCE CHARACTERISTICS:

10 (A) LENGTH: 53 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

20 (iv) ANTI-SENSE: NO

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 33:

25 ATAAATAGGT CTCTTTCATN NNNNNNNNNN NNNNNTCTAG AAATGTGTGA  
AAT 53

(2) INFORMATION FOR SEQ ID NO: 34:

30

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
35 (D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

-106-

- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- 5 (iv) ANTI-SENSE: NO
- 10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 34:
- TAACGAGGCG CAAAAA
- 16
- (2) INFORMATION FOR SEQ ID NO: 35:
- 15 (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 37 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- 20 (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- 25 (iv) ANTI-SENSE: NO
- 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 35:
- AGACACGTAC AAACCAATGA AAGAGACCTA TTTACTC
- 37

SUBSTITUTE SHEET (RULE 26)

-107-

## (2) INFORMATION FOR SEQ ID NO: 36:

- 5 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 33 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- 10 (ii) MOLECULE TYPE: cDNA  
(iii) HYPOTHETICAL: NO  
(iv) ANTI-SENSE: NO
- 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 36:

20 AAGCAAAGTC CCGCGAATGA AAGAGACCTA TTT

33

## (2) INFORMATION FOR SEQ ID NO: 37:

- 25 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 33 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- 30 (ii) MOLECULE TYPE: cDNA  
(iii) HYPOTHETICAL: NO  
(iv) ANTI-SENSE: NO
- 35

SUBSTITUTE SHEET (RULE 26)

-108-

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 37:

5 ACGTTTAAAC AGACACATGA AAGAGACCTA TTT

33

(2) INFORMATION FOR SEQ ID NO: 38:

(i) SEQUENCE CHARACTERISTICS:

10

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

15

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 38:

25

GGAACTCAAA GGCACC

16

(2) INFORMATION FOR SEQ ID NO: 39:

(i) SEQUENCE CHARACTERISTICS:

30

(A) LENGTH: 61 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

35

SUBSTITUTE SHEET (RULE 26)

-109-

- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- 5 (iv) ANTI-SENSE: NO
- 10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 39:
- CACACATTTTC TAGAAGACAC GTACAAACCA ATGAAAAAAG ACAGCTATCG 60  
CGATTGCAGT 61  
G
- 15 (2) INFORMATION FOR SEQ ID NO:40:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 87 base pairs
- 20 (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- 25 (iii) HYPOTHETICAL: NO
- (iv) ANTI-SENSE: NO
- 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 40:

SUBSTITUTE SHEET (RULE 26)

-110-

CACATTTCTA GAAAGCAAAG TCCCGCGAAT GAAAGAGACC TATTTATGAA  
AAAGACAGCT  
ATCGCGATTG CAGTGGCACT GGCTGGT

60

87

## 5 (2) INFORMATION FOR SEQ ID NO: 41:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 70 base pairs  
(B) TYPE: nucleic acid  
10 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: cDNA

## 15 (iii) HYPOTHETICAL: NO

## (iv) ANTI-SENSE: NO

20

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 41:

CACATTTCTA GAAAGCAAAG TCCCGCGAAT GAAAAAGACA GCTATCGCGA  
25 TTGCAGTGGC

60

ACTGGCTGGT

70

## (2) INFORMATION FOR SEQ ID NO: 42:

30

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 70 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
35 (D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

-111-

- (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- 5 (iv) ANTI-SENSE: NO
- 10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 42:
- CACATTTCTA GAACGTTTAA ACAGACACAT GAAAAAGACA GCTATCGCGA  
TTGCAGTGGC 60
- 15 ACTGGCTGGT 70
- (2) INFORMATION FOR SEQ ID NO: 43:
- (i) SEQUENCE CHARACTERISTICS:
- 20 (A) LENGTH: 70 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear
- 25 (ii) MOLECULE TYPE: cDNA
- (iii) HYPOTHETICAL: NO
- (iv) ANTI-SENSE: NO
- 30
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 43:
- 35

SUBSTITUTE SHEET (RULE 26)

-112-

CACATTTCTA GAGGAACTCA AAGGCACCAT GAAAAAGACA GCTATCGCGA  
TTGCAGTGGC

60

ACTGGCTGGT

70

-113-

## Claims:

1. A method for optimizing the production of polypeptide in a cell, comprising the steps of:
- 5 (a) identifying at least one regulatory region within a nucleic acid sequence to be mutagenized;
- (b) preparing a nucleic acid vector comprising said regulatory region and a nucleic acid sequence encoding at least one polypeptide regulated by said regulatory
- 10 region;
- (c) deleting said regulatory region from said vector;
- (d) producing a pool of random oligonucleotides;
- (e) using a polymerase chain reaction to
- 15 introduce at least one random oligonucleotide into the position previously occupied by said regulatory region in a plurality of vectors to generate a pool of mutagenized vectors;
- (f) introducing said mutagenized vectors into a
- 20 cell sample;
- (g) assaying for the expression of said polypeptide in said cell sample;
- (h) selecting cells exhibiting optimized polypeptide expression; and
- 25 (i) isolating optimized polypeptide from the cells of step (h).
2. The method of Claim 1, wherein said regulatory region is located within said nucleic acid sequence encoding polypeptide.
- 30 3. The method of Claim 1, wherein said regulatory region is in a translated portion of said nucleic acid sequence encoding polypeptide.
4. The method of Claim 1 wherein said regulatory region is selected from the group consisting of
- 35 a signal sequence, a ribosome binding site, a promoter

SUBSTITUTE SHEET (RULE 26)

-114-

sequence, a translational regulatory sequence, a transcription regulatory sequence, and a protein trafficking sequence.

5        5.    The method of Claim 1, wherein said nucleic acid sequence encodes an antibiotic resistant gene and the selection step additionally consists essentially of growing said cell sample in the presence of an antibiotic.

10       6.    The method of Claim 1, wherein said nucleic acid sequence encoding polypeptide encodes a selectable marker.

7.    The method of Claim 1, wherein said nucleic acid sequence encoding polypeptide encodes a fusion protein.

15       8.    The method of Claim 1, wherein said nucleic acid sequence encoding polypeptide is derived from a eukaryotic cell and said cell samples is prokaryotic.

9.    The method of Claim 1, wherein said random oligonucleotides are biased.

20       10.   The method of Claim 1, additionally comprising the steps of introducing said vector comprising said regulatory region and a nucleic acid sequence encoding at least one polypeptide regulated by said regulatory region into a second cell sample, expressing said polypeptide encoded by said vector in said second cell sample, and measuring the level of polypeptide expression in said second cell sample.

25       11.   The method of Claim 10, additionally comprising the step of selecting cells from said first cell sample that exhibit optimized polypeptide expression relative to the measuring step of Claim 10.

30       12.   The method of Claim 11, additionally comprising the step of determining the nucleic acid sequence of the mutagenized regulatory region producing optimized polypeptide expression.

-115-

13. The method of Claim 11, wherein said optimized polypeptide expression is a level of polypeptide expression greater than or equal to the level of polypeptide expression obtained from the measuring step of Claim 10.

14. A method for creating and isolating novel signal sequences, comprising the steps of:

- (a) identifying a signal sequence region within a nucleic acid sequence encoding a polypeptide;
- 10 (b) preparing a nucleic acid vector comprising said nucleic acid sequence encoding the polypeptide;
- (c) introducing said vector into a first cell sample and expressing said polypeptide in said first cell sample;
- 15 (d) measuring the level of polypeptide expression in said first cell sample;
- (e) deleting said signal sequence region from said vector;
- (f) producing a pool of random oligonucleotides;
- 20 (g) using a polymerase chain reaction to introduce at least one of the random oligonucleotides into the position previously occupied by said signal sequence in a plurality of vectors to generate a pool of mutagenized vectors;
- 25 (h) introducing said mutagenized vectors into a second cell sample;
- (i) assaying for the expression of said polypeptide in said second cell sample;
- (j) selecting cells exhibiting optimized
- 30 polypeptide expression relative to step (d); and
- (k) determining the nucleic acid sequence of the regulatory region located in the mutagenized vector contained in the cells of Step (j).

15. The method of Claim 14, wherein said signal sequence region comprises an entire signal sequence.

-116-

16. The method of Claim 14, wherein said nucleic acid vector encodes an antibiotic resistant gene and the selection step additionally consists essentially of growing said cell sample in the presence of an antibiotic.

5 17. The method of Claim 14, wherein said nucleic acid sequence encoding polypeptide encodes a selectable marker.

18. The method of Claim 14, wherein said polymerase chain reaction is the enzymatic inverse  
10 polymerase chain reaction.

19. The method of Claim 14, wherein said nucleic acid sequence encoding polypeptide encodes a fusion protein.

20. The method of Claim 16, wherein a portion of  
15 said fusion protein is derived from  $\beta$ -lactamase.

21. The method of Claim 20, wherein a portion of said fusion protein is derived from an antibody.

22. The method of Claim 15, wherein said nucleic acid sequence encoding polypeptide is derived from a  
20 eukaryotic cell and said first and second cell samples are prokaryotic.

23. The method of Claim 22, wherein said cells are *Escherichia coli*.

24. The method of Claim 15, wherein said random  
25 oligonucleotides are biased.

25. The method of Claim 24, wherein said random oligonucleotides contain at least one positively charged amino acid at the N-terminus, a stretch of at least 8 hydrophobic amino acids and a small amino acid such as  
30 alanine, glycine, or valine positioned at the C-terminus.

26. The method of Claim 15, wherein said optimized level of polypeptide expression is a level greater than or equal to the level of expression of said polypeptide in step (d).

SUBSTITUTE SHEET (RULE 26)

-117-

27. A method for optimizing polypeptide expression in a cell by performing random mutagenesis on a regulatory region regulating polypeptide expression from a nucleic acid vector wherein the improvement comprises  
5 deleting a region of nucleic acid to be mutagenized, isolating said vector containing said deletion and replacing said region with random nucleic acid sequences.

28. A method for creating and isolating novel ribosome binding sites, comprising the steps of:

10 (a) identifying a ribosome binding site within a nucleic acid sequence;

(b) preparing a nucleic acid vector comprising said nucleic acid sequence containing a ribosome binding site operably linked to a nucleic acid sequence encoding a  
15 polypeptide;

(c) introducing said vector into a first cell sample and expressing said polypeptide in said first cell sample;

(d) measuring the level of polypeptide  
20 expression in said first cell sample;

(e) deleting said nucleic acid sequence containing said ribosome binding site from said vector;

(f) producing a pool of random oligonucleotides;

(g) using a polymerase chain reaction to  
25 introduce at least one of the random oligonucleotides into the position previously occupied by said nucleic acid sequence containing said ribosome binding site in a plurality of vectors to generate a pool of mutagenized vectors;

30 (h) introducing said mutagenized vectors into a second cell sample;

(i) assaying for the expression of said polypeptide in said second cell sample;

(j) selecting cells exhibiting optimized  
35 polypeptide expression relative to step (d); and

SUBSTITUTE SHEET (RULE 26)

-118-

(k) determining the nucleic acid sequence containing said ribosome binding site located in the mutagenized vector introduced into the cells of step (j).

29. The method of Claim 28, wherein said nucleic  
5 sequence encodes an antibiotic resistance gene and said selection step consists essentially of growing said cell samples in the presence of an antibiotic.

30. The method of Claim 28, wherein said nucleic  
10 acid sequence encoding polypeptide encodes a selectable marker.

31. The method of Claim 30, wherein said nucleic  
acid sequence encoding polypeptide is a fusion protein.

32. The method of Claim 30, wherein a portion of  
said fusion protein is derived from an antibody.

15 33. The method of Claim 31, wherein a portion of  
said fusion protein is derived from  $\beta$ -lactamase.

34. The method of Claim 28, wherein said nucleic  
acid sequence encoding polypeptide is derived from a  
eukaryotic cell and said first and second cell samples are  
20 prokaryotic.

35. The method of Claim 31, wherein said cells  
are *Escherichia coli*.

36. The method of Claim 28, wherein said random  
oligonucleotides are biased.

25 37. The method of Claim 28, wherein said  
polymerase chain reaction is the enzymatic inverse  
polymerase chain reaction.

38. The method of Claim 28, wherein said  
optimized level of polypeptide expression is a level  
30 greater than or equal to the level of expression of said  
polypeptide in step (d).

39. A protein signal sequence corresponding to  
SEQ ID NO: 13 and to protein signal sequences containing at  
least contiguous amino acid 10-mers thereof.

-119-

40. A protein signal sequence corresponding to SEQ ID NO: 16 and to protein signal sequences containing at least contiguous amino acid 10-mers thereof.

5 41. A protein signal sequence corresponding to SEQ ID NO: 14 and to protein signal sequences containing at least contiguous amino acid 10-mers thereof.

42. Isolated nucleic acid containing a ribosome binding site sequence identified as SEQ ID NO: 35 and to ribosome binding sites containing at least contiguous 5-  
10 mers thereof.

43. Isolated nucleic acid containing a ribosome binding site sequence identified as SEQ ID NO: 37 and to ribosome binding sites containing at least contiguous 5-mers thereof.

15 44. Isolated nucleic acid containing a ribosome binding site sequence identified as SEQ ID NO: 38 and to ribosome binding sites containing at least contiguous 5-mers thereof.

45. A method for creating and isolating novel  
20 signal sequences, comprising the steps of:

(a) identifying a signal sequence region within a nucleic acid sequence encoding a polypeptide;

(b) preparing a nucleic acid vector comprising said nucleic acid sequence encoding a polypeptide, wherein  
25 said polypeptide is a fusion protein having a C-terminus corresponding to  $\beta$ -lactamase;

(c) deleting said signal sequence region from said vector;

(d) producing a pool of random oligonucleotides  
30 suitable for an enzymatic inverse polymerase chain reaction;

(e) using an enzymatic inverse polymerase chain reaction to introduce said random oligonucleotides into the position in said vector previously occupied by said signal  
35 sequence to generate a pool of mutagenized vectors;

SUBSTITUTE SHEET (RULE 26)

-120-

(f) introducing said mutagenized vectors into a first cell sample of *E. coli*;

(g) assaying for the expression of said polypeptide in said first cell sample in the presence of  
5 ampicillin; and

(h) selecting cells exhibiting optimized polypeptide expression.

46. The method of Claim 45, wherein said signal sequence region comprises an entire signal sequence.

10 47. The method of Claim 45, additionally comprising the steps of introducing said vector comprising said nucleic acid sequence encoding a polypeptide, wherein said polypeptide is a fusion protein having a C-terminus corresponding to  $\beta$ -lactamase into a second cell sample of,  
15 *E. coli* expressing said polypeptide in said second cell sample and measuring the level of polypeptide expression in said second cell sample.

48. The method of Claim 47, additionally comprising the step of selecting cells from said first cell  
20 sample that exhibit optimized polypeptide expression relative to the level of polypeptide expression observed in the measuring step of Claim 49.

49. The method of Claim 48, wherein said optimized polypeptide expression is a level of polypeptide  
25 expression greater than or equal to the level of polypeptide expression obtained from the measuring step of Claim 47.

50. The method of Claim 46 or 48 additionally comprising the step of determining the nucleic acid  
30 sequence of said regulatory region contained in the mutagenized vector isolated from cells identified by said selecting step.

51. The method of Claim 46, wherein the concentration of ampicillin in step (i) is at least 30  
35  $\mu\text{g/ml}$ .

-121-

52. The method of Claim 46, wherein said nucleic acid sequence encoding a polypeptide encodes a single-chain antibody.

53. The method of Claim 46, wherein said random  
5 oligonucleotides are biased.

54. The method of Claim 53, wherein said random oligonucleotides contain at least one positively charged amino acid at the N-terminus, a stretch of at least 8 hydrophobic amino acids and a small amino acid such as  
10 alanine, glycine or valine positioned at the C-terminus.

55. A method for identifying a novel nucleic acid sequence encoding a protein trafficking signal that directs a polypeptide to a desired location in a cell, comprising the steps of:

15 (a) identifying a region of nucleic acid containing at least one protein trafficking signal to be mutagenized;

(b) preparing a nucleic acid vector comprising said protein trafficking signal sequence to be mutagenized  
20 and a nucleic acid sequence encoding at least one polypeptide;

(c) deleting said protein trafficking signal sequence from said vector;

(d) producing a pool of random oligonucleotides;

25 (e) using a polymerase chain reaction to introduce at least one of the random oligonucleotides into the position previously occupied by said protein trafficking signal in a plurality of vectors to generate a pool of mutagenized vectors;

30 (f) introducing said mutagenized vectors into a cell sample;

(g) assaying for the location of said polypeptide in said cell sample;

35 (h) selecting cells expressing said polypeptide in the desired cell location; and

SUBSTITUTE SHEET (RULE 26)

-122-

(i) determining the nucleic acid sequence of the protein trafficking signal contained in the mutagenized vector located in the cells of Step (h).

5 56. The method of Claim 55, wherein said protein trafficking sequence to be mutagenized is located in said nucleic acid sequence encoding said polypeptide.

57. The method of Claim 55, wherein said polymerase chain reaction is the enzymatic inverse polymerase chain reaction.

10 58. The method of Claim 55, wherein said desired cell location is extracellular.

59. The method of Claim 55, wherein said nucleic sequence encodes an antibiotic resistant gene and said selection step consists essentially of growing said cell  
15 sample in the presence of an antibiotic.

60. The method of Claim 55, wherein said polypeptide is a selectable marker.

61. The method of Claim 55, wherein said polypeptide is a fusion protein.

20 62. The method of Claim 55, wherein said polypeptide is derived from a eukaryotic cell and said cell sample is prokaryotic.

63. The method of Claim 55, wherein said random oligonucleotides are biased.

25 64. A method for creating and isolating novel regulatory sequences for optimizing the expression of a recombinant polypeptide in prokaryotic cells comprising the steps of:

30 (a) identifying at least one regulatory region within a nucleic acid sequence to be mutagenized;

(b) preparing a nucleic acid vector suitable for expressing polypeptide in a prokaryotic cell, said nucleic acid vector comprising said regulatory region and a nucleic acid sequence encoding at least one polypeptide operably  
35 linked to said regulatory region;

**SUBSTITUTE SHEET (RULE 26)**

-123-

- (c) introducing said vector into a first prokaryotic cell sample and expressing said polypeptide encoded by said vector in said first cell sample;
- (d) measuring the level of polypeptide expression in said first prokaryotic cell sample;
- 5 (e) deleting said regulatory region from said vector;
- (f) producing a pool of random oligonucleotides;
- (g) using a polymerase chain reaction to
- 10 introduce at least one random oligonucleotide into said position previously occupied by said regulatory region in a plurality of vectors to generate a pool of mutagenized vectors;
- (h) introducing said mutagenized vectors into a
- 15 first prokaryotic cell sample;
- (i) assaying for the expression of the polypeptide in said second prokaryotic cell sample;
- (j) selecting cells exhibiting optimized polypeptide expression relative to step (d); and
- 20 (k) determining the nucleic acid sequence of said mutagenized regulatory region located within the vector introduced into the cells of step (j).
65. The method of Claim 64, wherein said regulatory region comprises an entire regulatory sequence.
- 25 66. The method of Claim 65, wherein said regulatory region is selected from the group consisting of a signal sequence, a ribosome binding site, a promoter sequence, a translational regulatory sequence, a transcription regulatory sequence and a protein trafficking
- 30 sequence.
67. The method of Claim 65, wherein said nucleic acid sequence encoding polypeptide encodes an antibiotic resistant gene and the selection step additionally consists of growing said second prokaryotic cell sample in the
- 35 presence of an antibiotic.

-124-

68. The method of Claim 65, wherein said nucleic acid sequence encoding polypeptide encodes a selectable marker.

5 69. The method of Claim 65, wherein said nucleic acid sequence encoding polypeptide encodes a fusion protein.

70. The method of Claim 65, wherein said random oligonucleotides are biased.

10 71. The method of Claim 65, wherein said nucleic acid sequence encoding polypeptide encodes a polypeptide derived from a eukaryotic cell.

72. The method of Claim 65, wherein said first and second cell samples are derived from the same cell type.

1 / 22

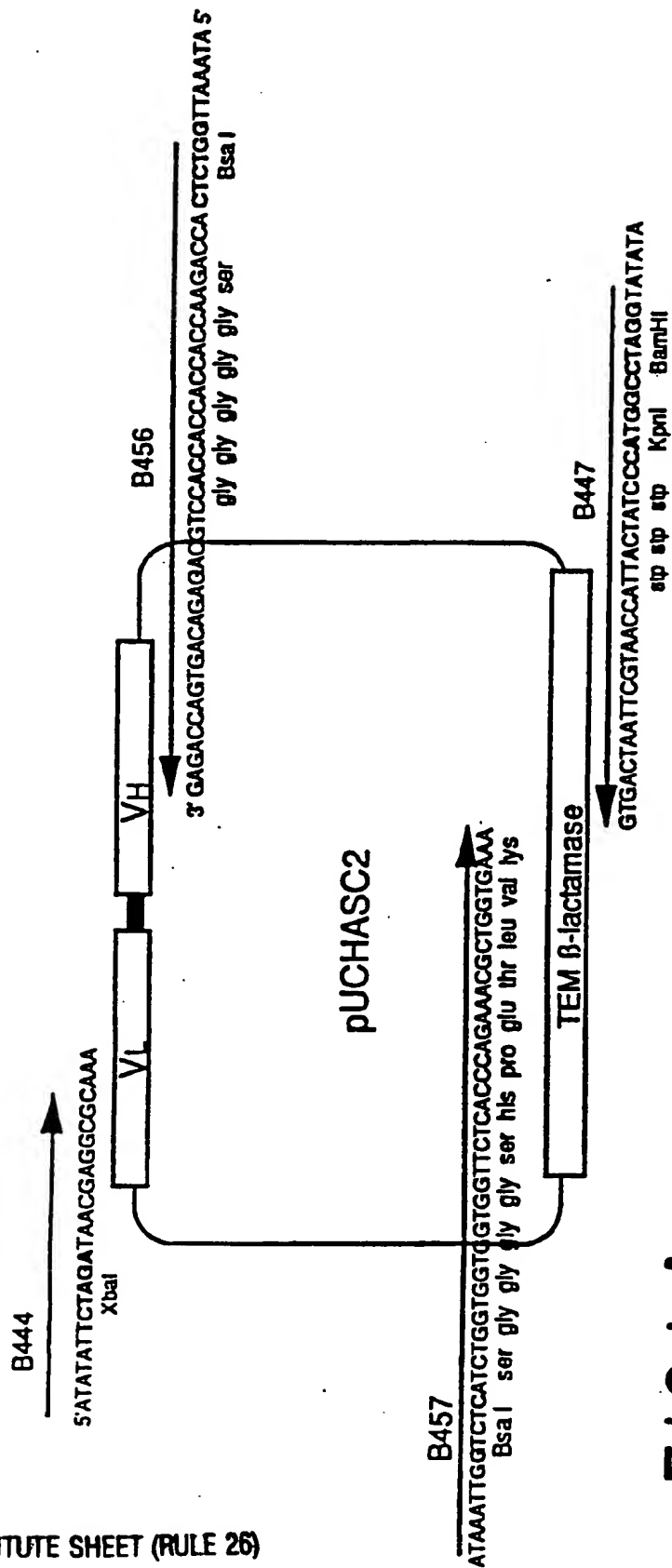


FIG. 1 A

2 / 22

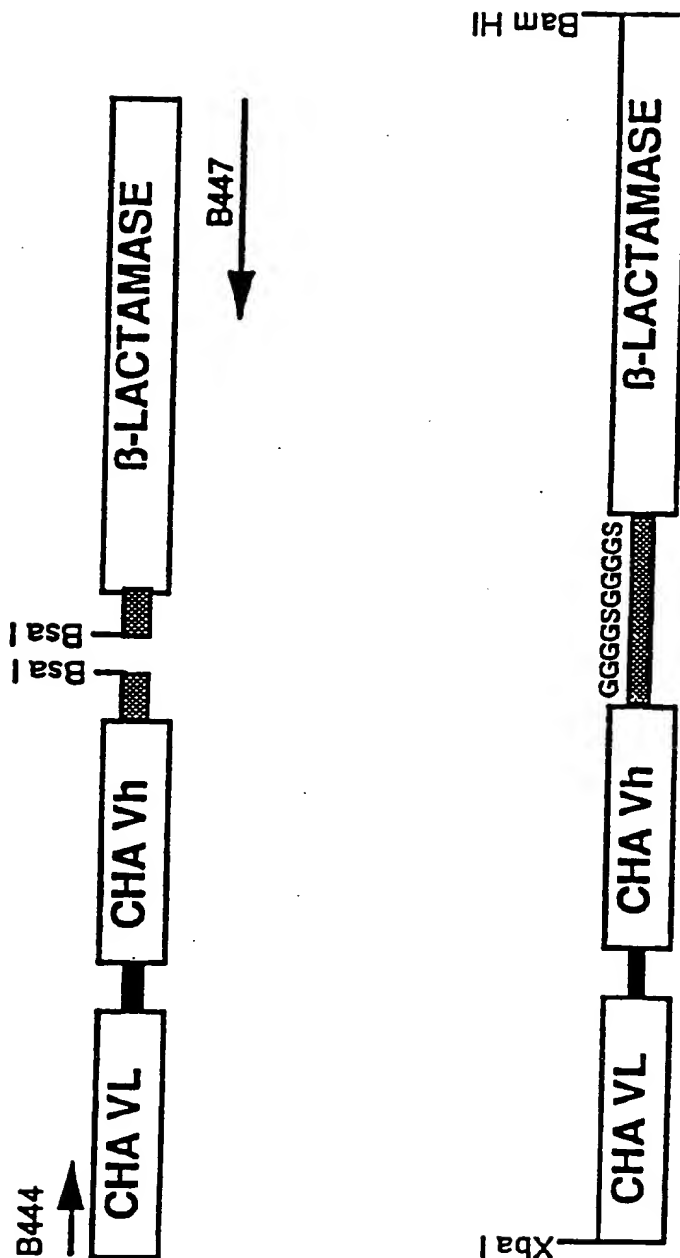


FIG. 1 B

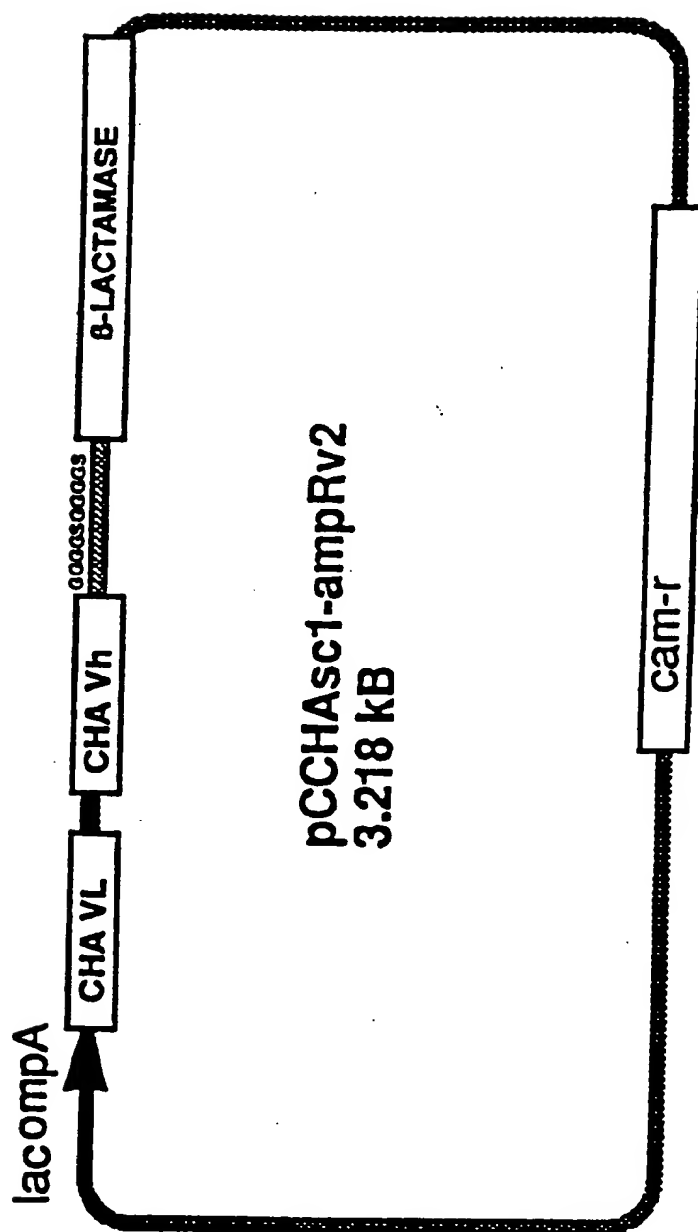


FIG.1C

4 / 22

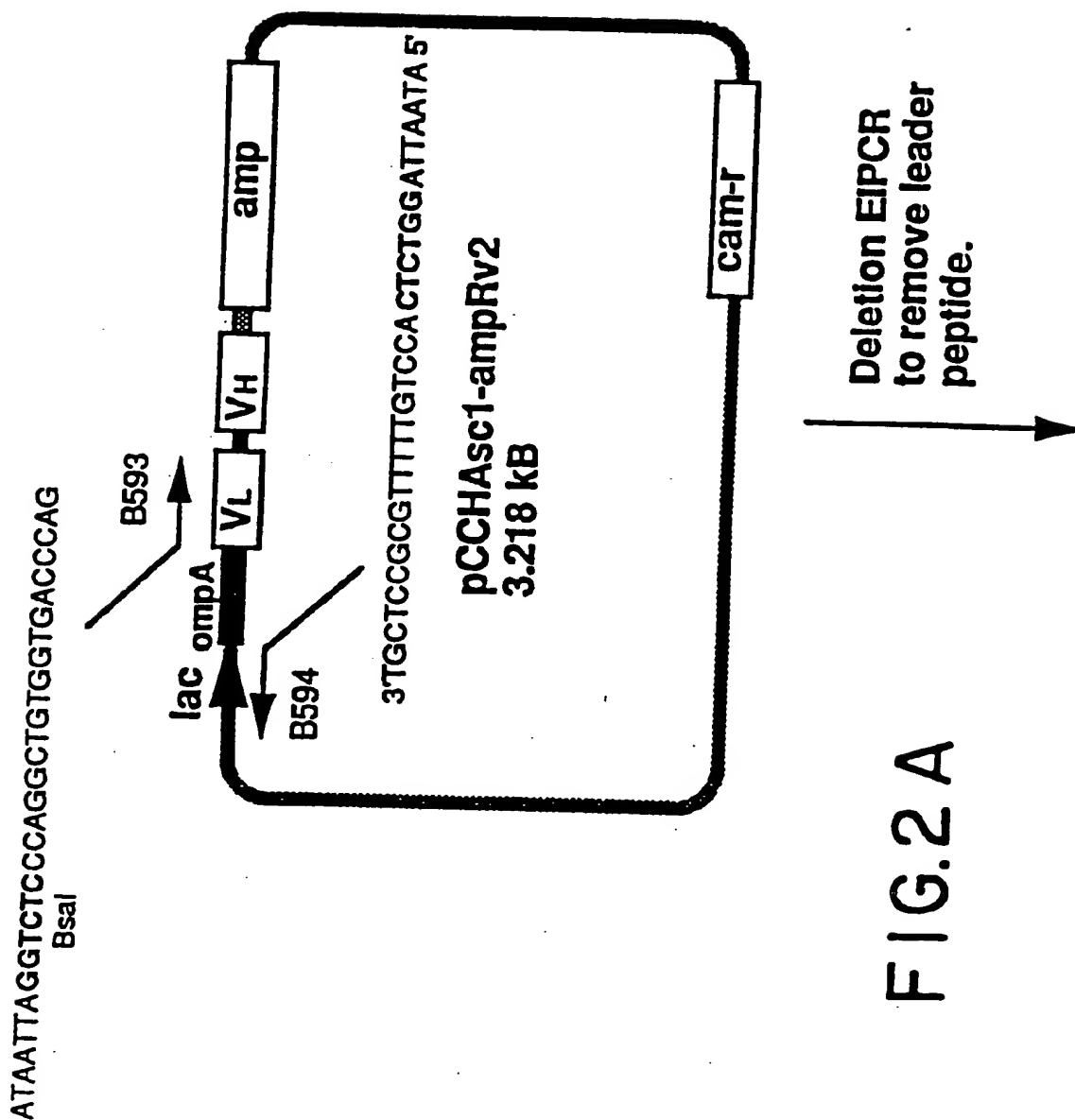
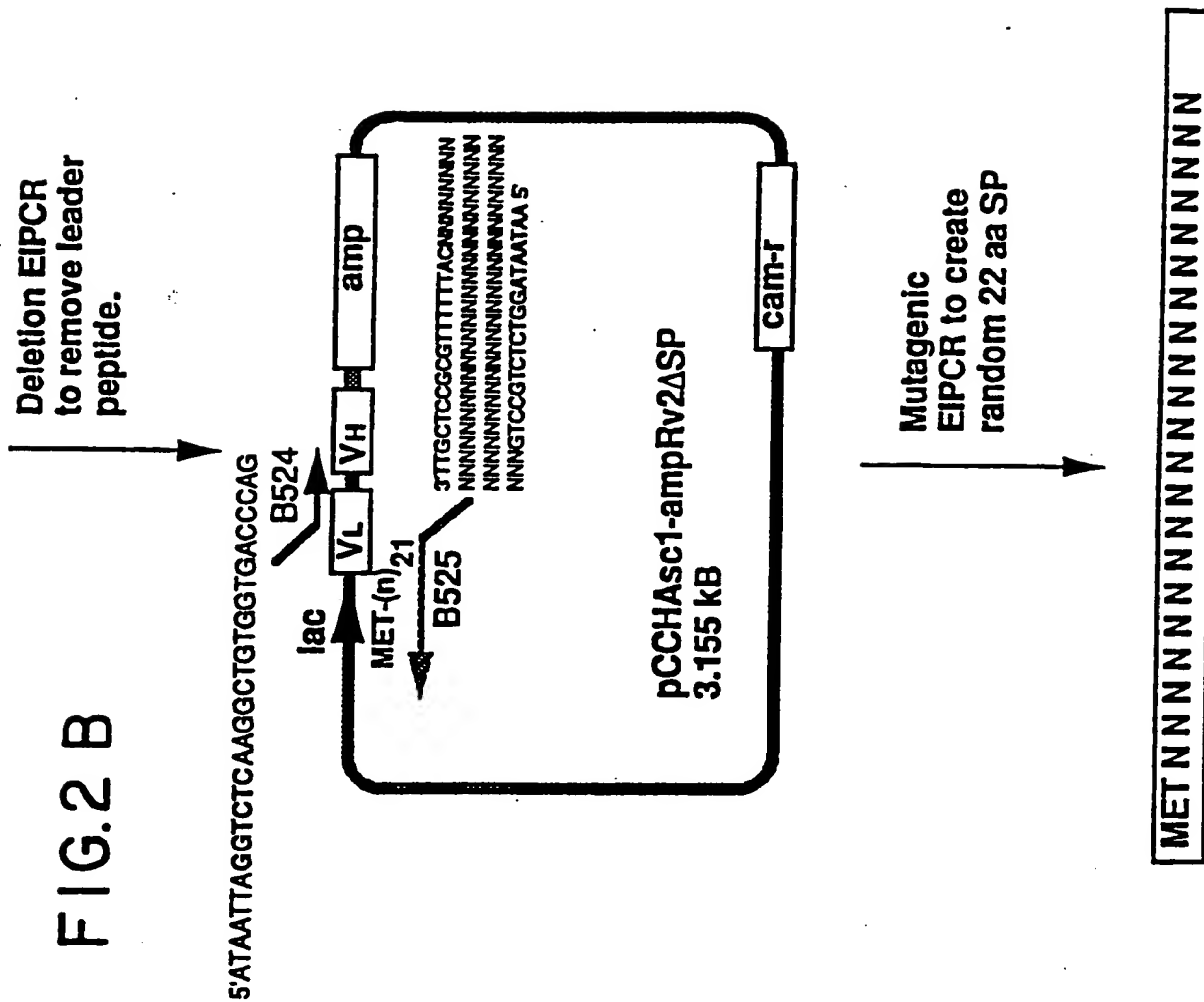


FIG.2 A

5 / 22



6 / 22

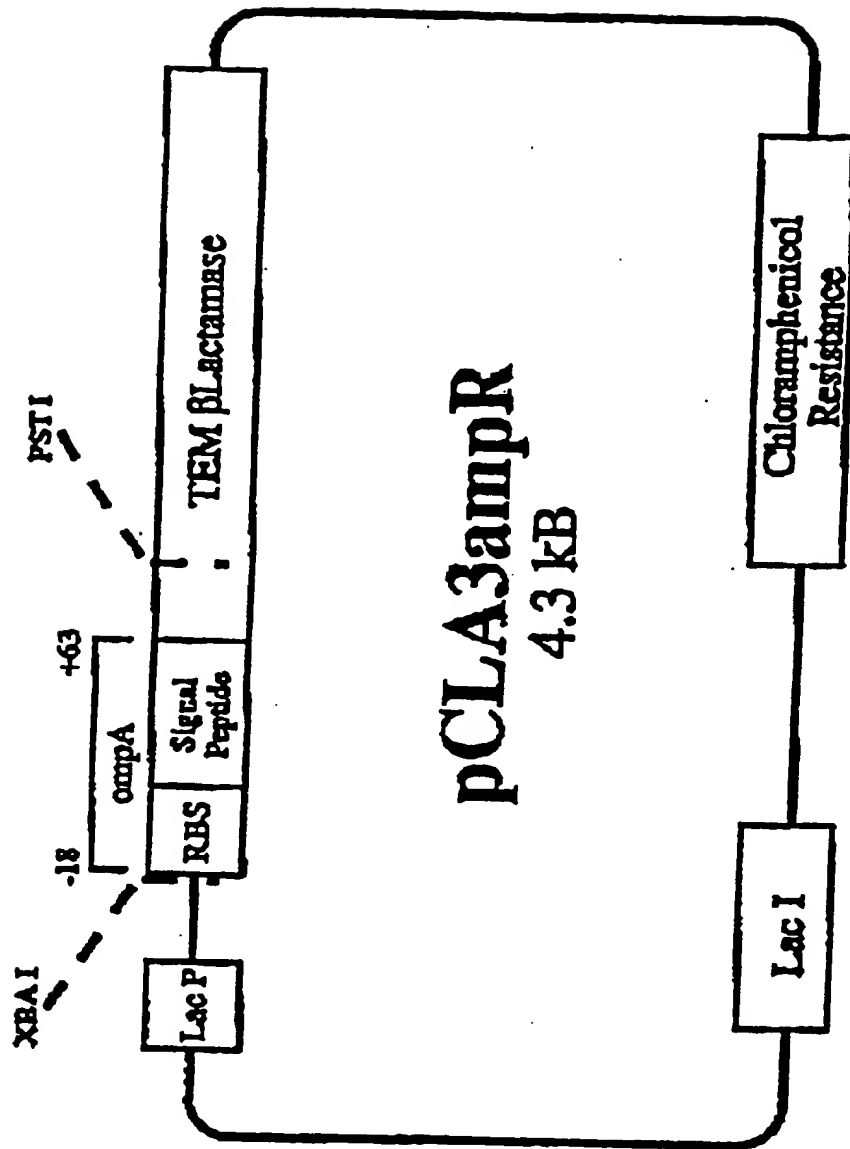


FIG. 3

7 / 22

# Wildtype OMPA

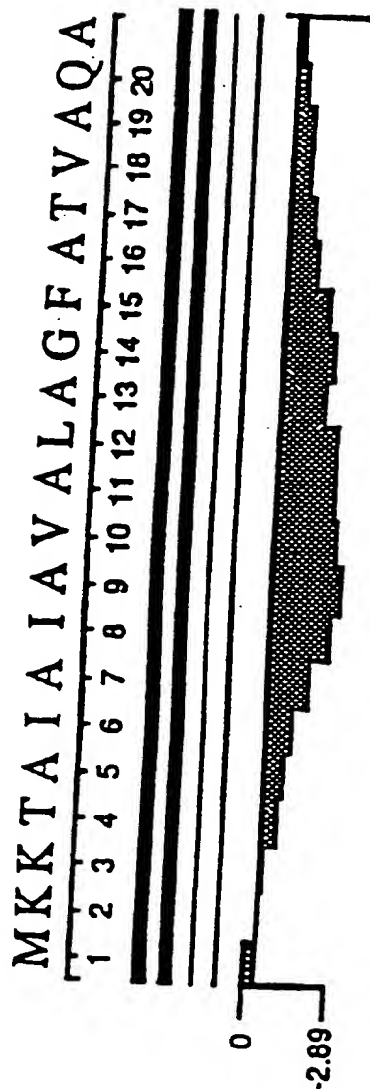


FIG. 4 A

8 / 22

E-10

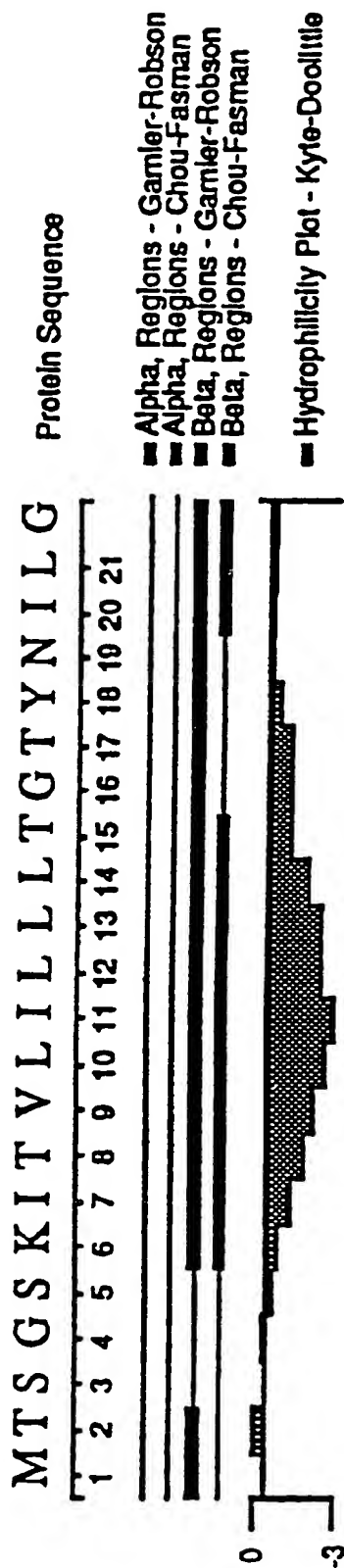


FIG. 4B

9 / 22

E-5

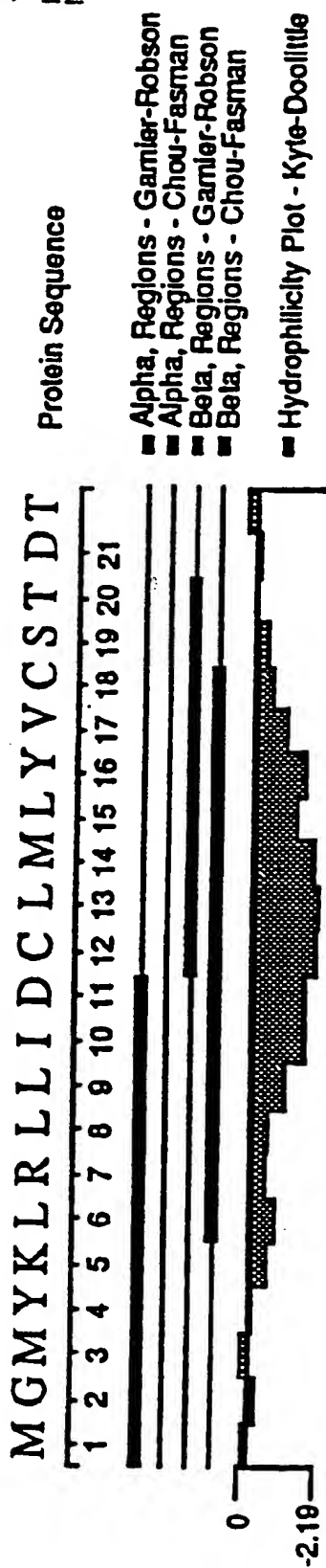


FIG. 4C

10/22

# PELB

M K Y L L P T A A A G L L L L A A Q P A M

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19, 20

Protein Sequence

- Alpha, Regions - Garnier-Robson
- Alpha, Regions - Chou-Fasman
- Beta, Regions - Garnier-Robson
- Beta, Regions - Chou-Fasman

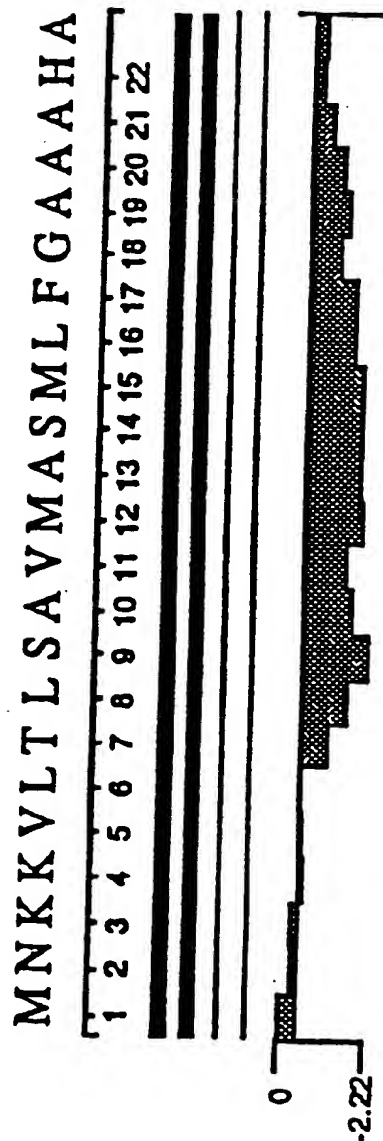
- Hydrophilicity Plot - Kyte-Doolittle



FIG. 4D

11/22

# DGAL



- Alpha, Regions - Gamler-Robson
- Alpha, Regions - Chou-Fasman
- Beta, Regions - Gamler-Robson
- Beta, Regions - Chou-Fasman
- Hydropathicity Plot - Kyte-Doolittle

FIG. 4E

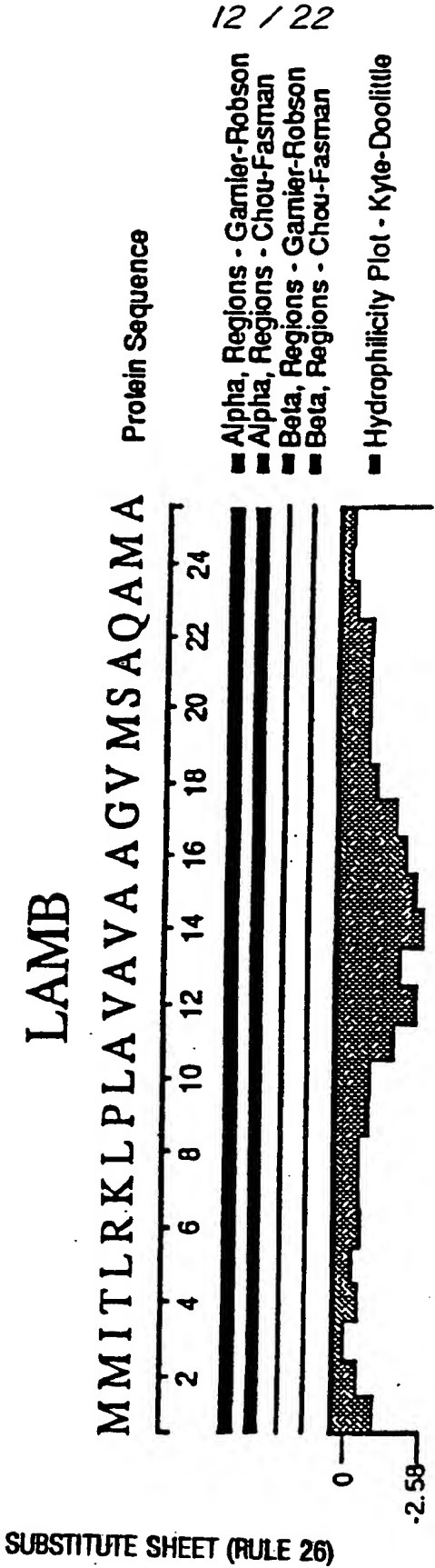


FIG. 4F

13 / 22

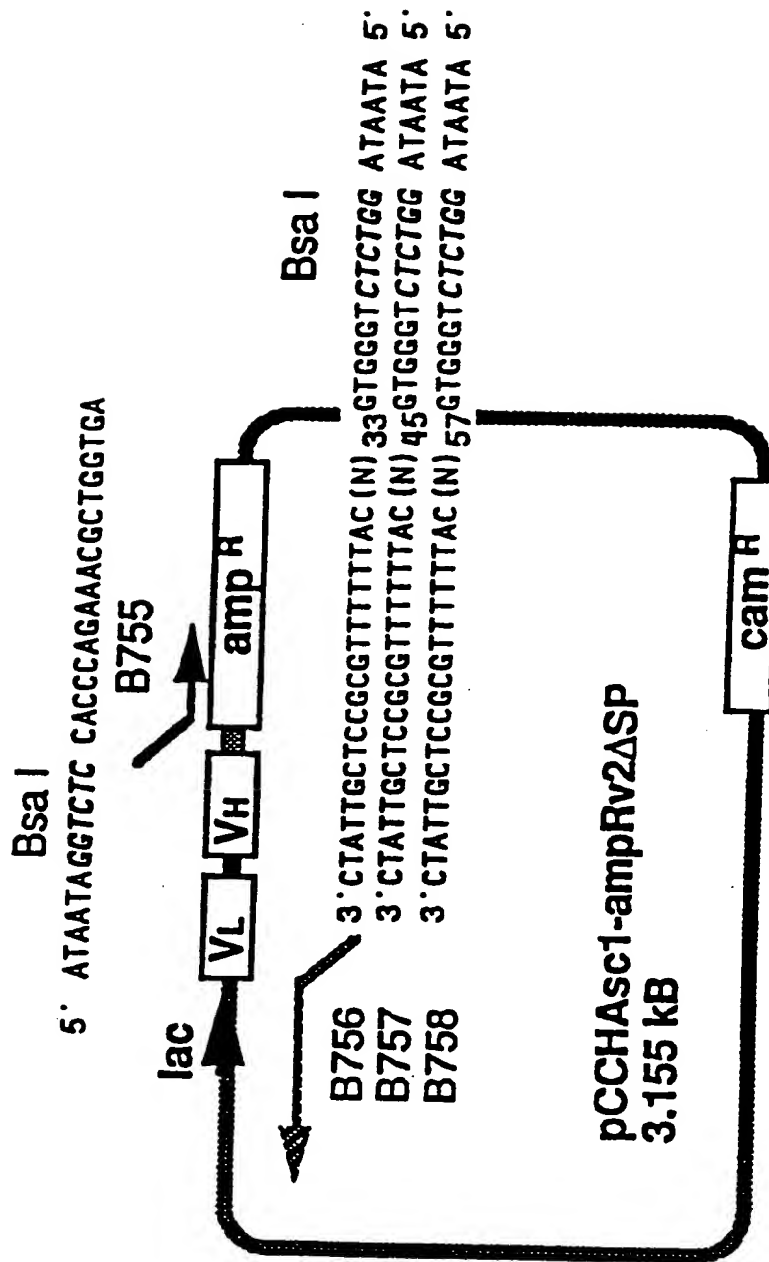
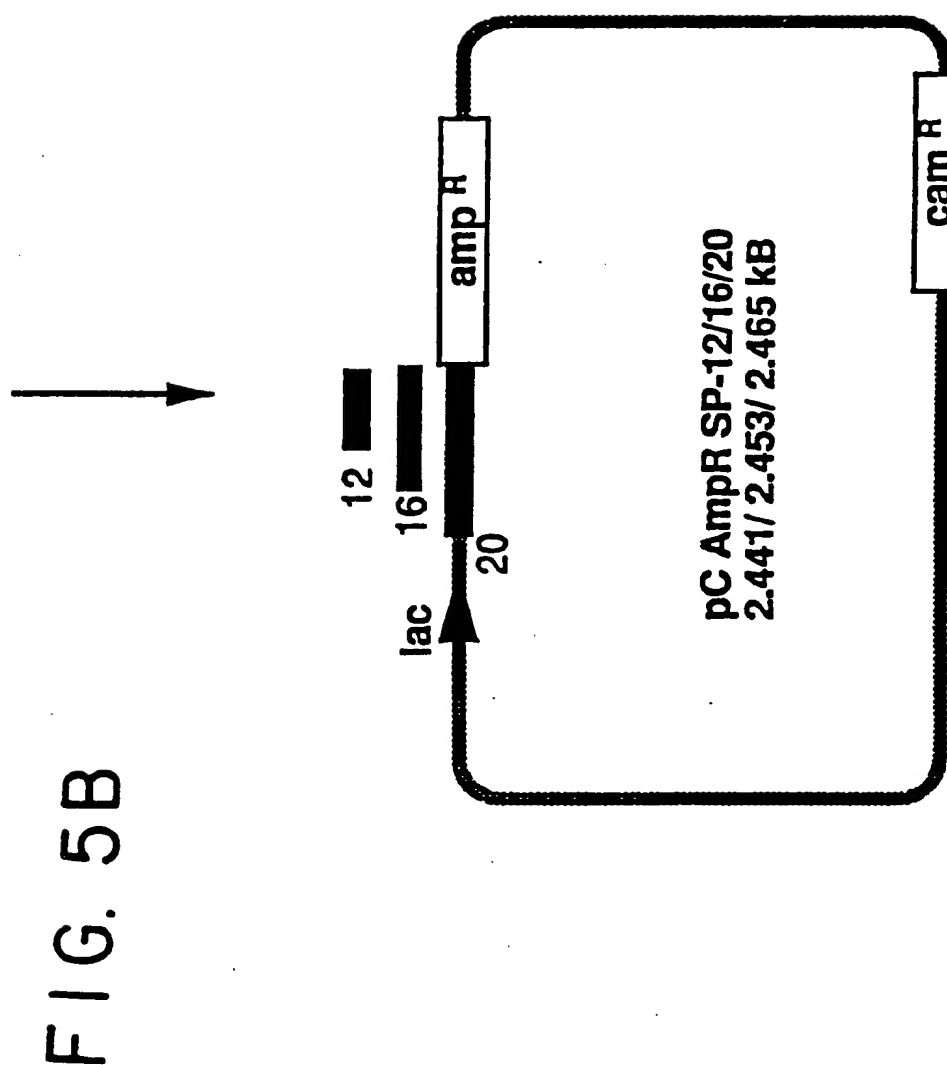


FIG. 5A

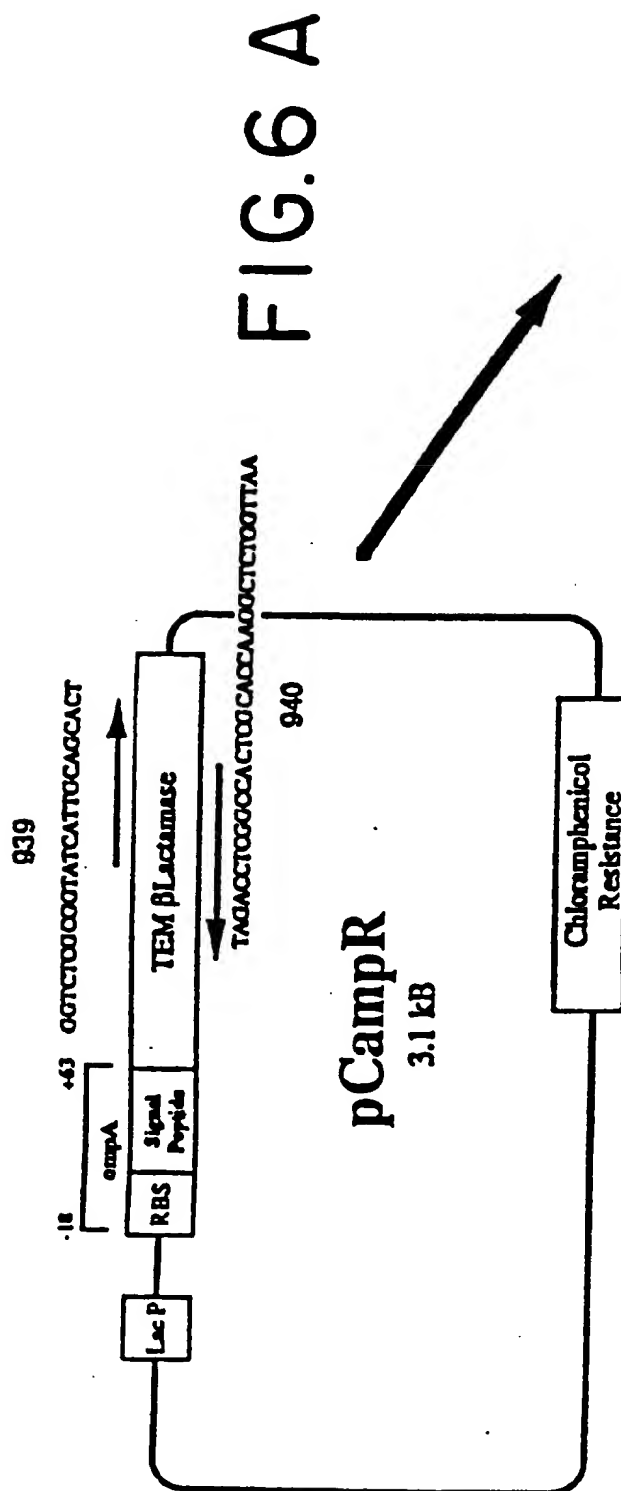
14 / 22



Plasmid Libraries containing random leader peptide of 12, 16 or 20 amino acids.  
The random signal peptides are used to direct  $\beta$ -lactamase to the periplasm of

E. coli

15 / 22



16 / 22

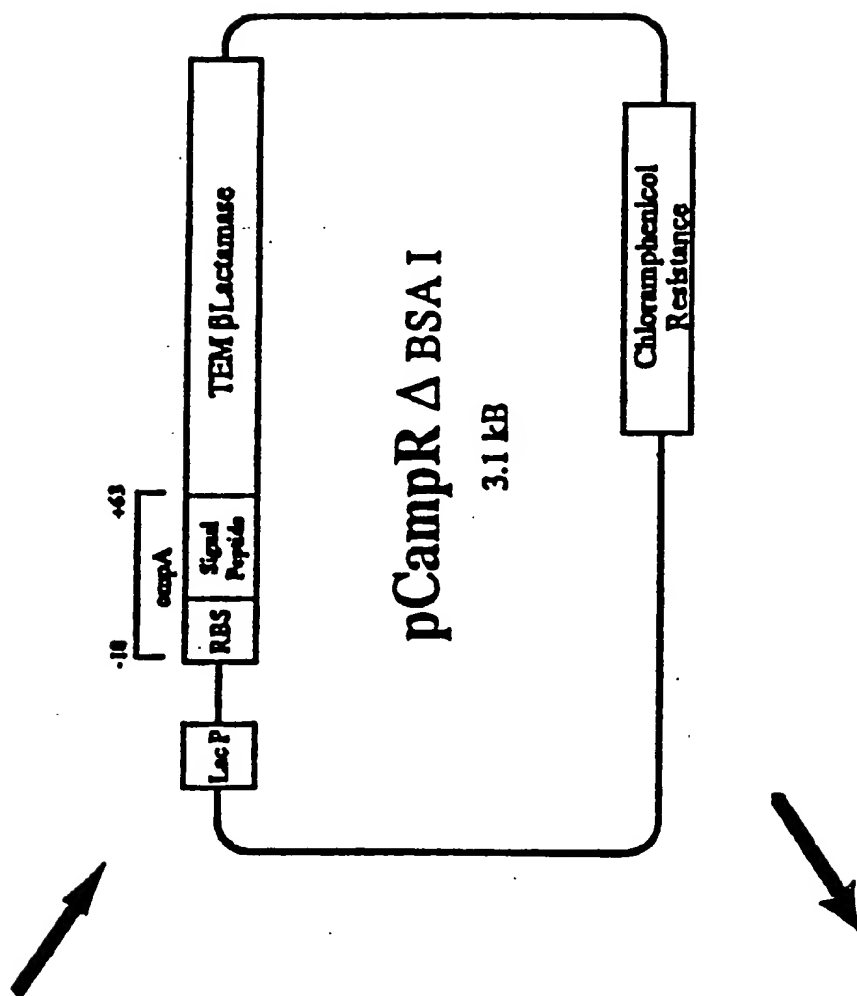
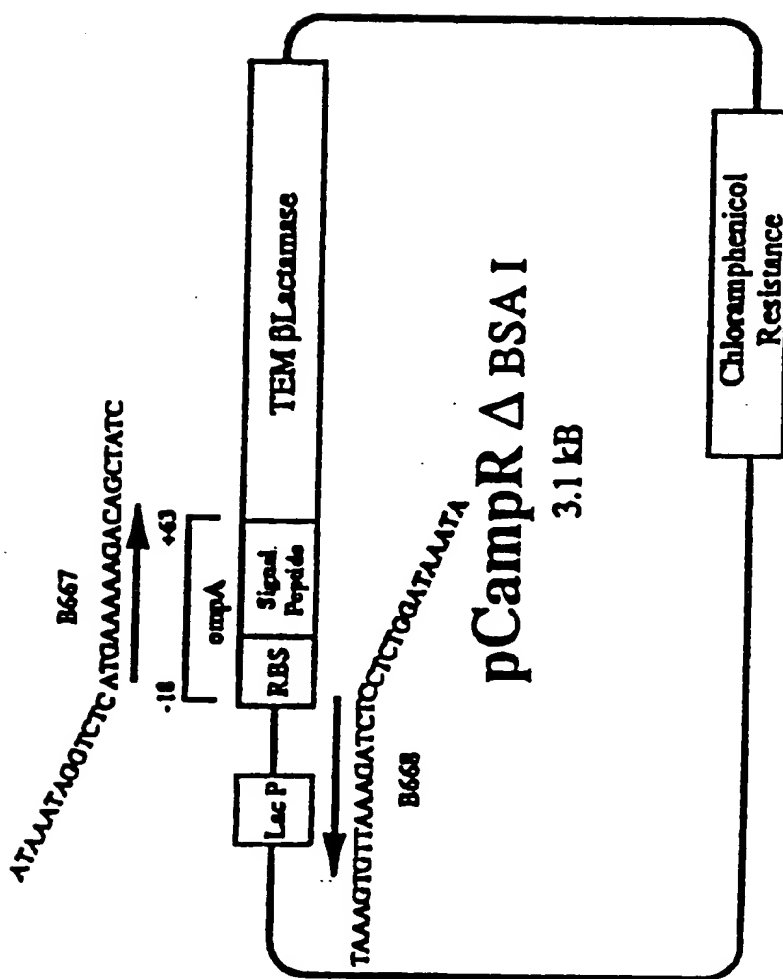


FIG. 6B

17/22

FIG. 6C



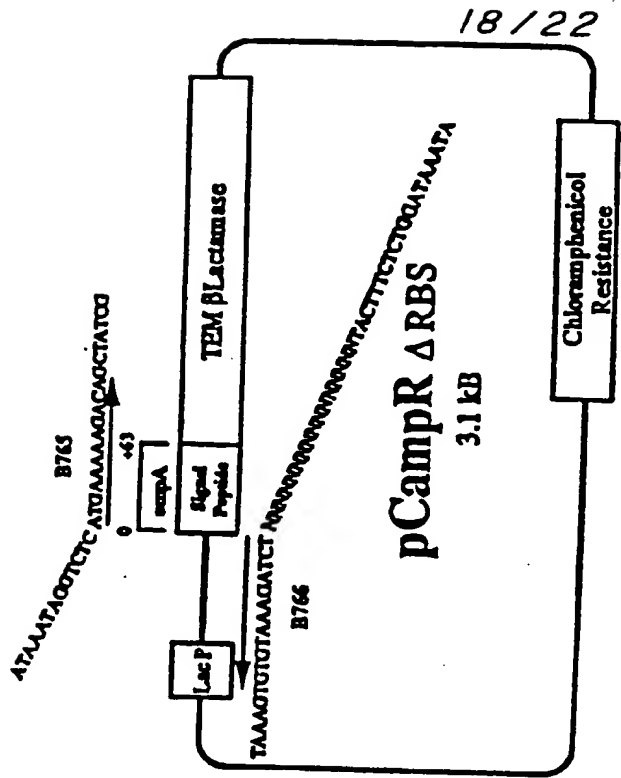


FIG.6D

Xba I      16 Base Random TRBS      MET

TCTAGA NNNNNNNNNNNNNNNN ATG

19 / 22

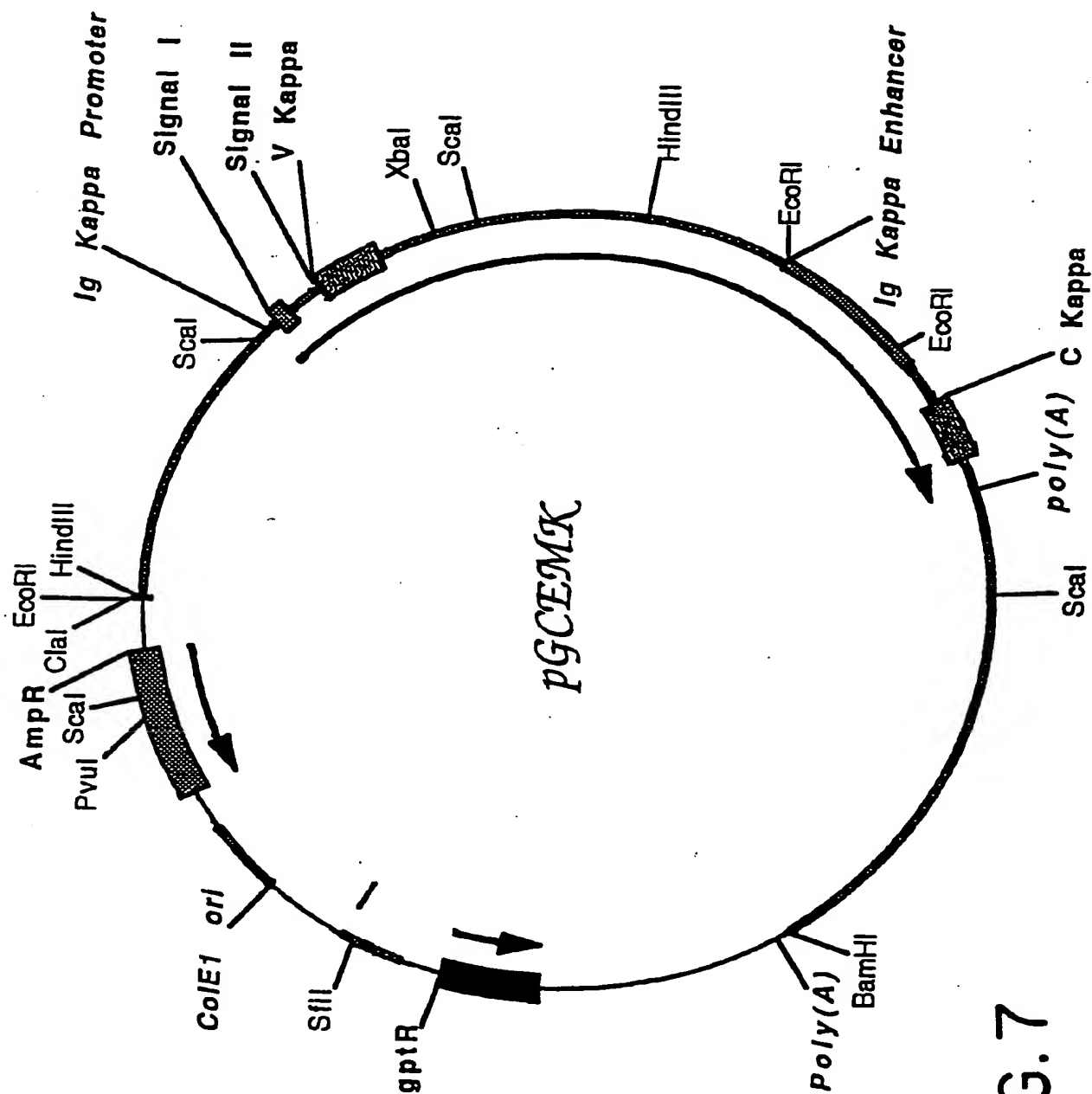


FIG. 7

20 / 22

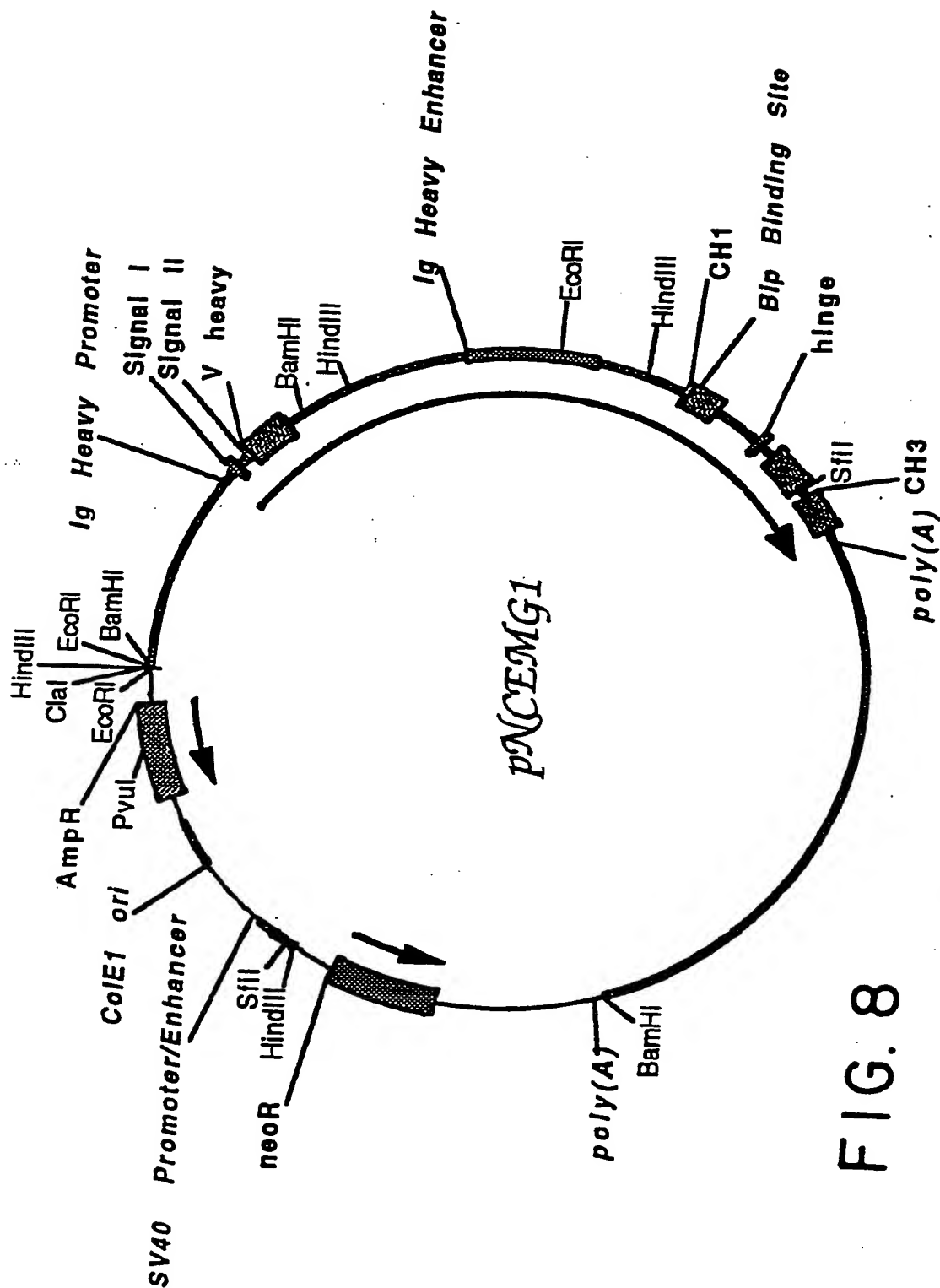


FIG. 8

21 / 22

RBS PrimersRBS 1 B949 (SEQ ID NO: 39)

5' cacacatttctagaAGACACGTACAAACCAatgaaaaagacagctatcgcgattgcagtg 3'

RBS 2 long B849 (SEQ ID NO: 40)5' cacatttctagaAAGCAAAGTCCCGGAAATGAAGAGACCTATTatgaaaaagacag  
ctatcgcgattgcagtggcactggctgggt 3'Upstream  
PrimersRBS 2 short B850 (SEQ ID NO: 41)

5' cacatttctagaAAGCAAAGTCCCGGAAatgaaaaagacagctatcgcgattgcagtggcactggctgggt 3'

RBS 8 short B844 (SEQ ID NO: 42)

5' cacatttctagaACGTTTAAACAGACACatgaaaaagacagctatcgcgattgcagtggcactggctgggt 3'

RBS 12 B842 (SEQ ID NO: 43)

5' cacatttctagaGGAACCTCAAAGGCACCAatgaaaaagacagctatcgcgattgcagtggcactggctgggt 3'

Downstream

940 (SEQ ID NO: 4)

Primer

5' aattggctcggaaaccacgctaccggctccagat 3'

FIG. 9A

Signal Peptide PrimersE10 B946 (SEQ ID NO: 25)

5' taattattctaga**ATG**ACTTCTGGGAGCAAAATAACTGTTCTTATACTTCTTTG  
 ACTGGTACATACAATATATTAGGGgaattccgggtcaccagaaacgctcgggaaagtaaaaga 3'

E5 B947 (SEQ ID NO: 24)

5' taattattctaga**ATG**GGGATGTACAAGCTGAGACTACTGATAGATTGCCCTAAT  
 GCTTATGTTGTTCAACCGATACCGgaattccgggtcaccagaaacgctcgggaaagtaaaaga 3'

Upstream  
Primers940 (SEQ ID NO: 4)

5' aatggctcgcgaaccacgctcaccggctccagat 3'

Downstream  
Primer

Note: Letters in bold indicate the sequence taken from the RBS  
 library clones (between the Xba I site and the ATG start codon  
 of the signal peptide).

FIG. 9B

22 / 22

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US94/04651**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(5) : C12N 15/63, 15/09; C12P 21/00

US CL : 435/69.1, 69.7, 172.3, 252.3, 320.1; 536/23.4, 24.1

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/69.1, 69.7, 172.3, 252.3, 320.1; 536/23.4, 24.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched  
NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, STN/MEDLINE

search terms: random?, promoter#, express?.

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US, A, 4,683,195 (MULLIS ET AL.) 28 July 1987, line 15 of column 29 to line 18 of column 30.	1-72
Y	US, A, 5,096,815 (LADNER ET AL) 17 March 1992, see entire document.	1-72
Y	SCIENCE Vol. 235, issued 16 January 1987, C.A. Kaiser et al., "Many Random Sequences Functionally Replace the Secretion Signal Sequence of Yeast Invertase", pages 312-317, see entire document.	1-27, 39-41, 45-63

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* documents defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E* earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L* documents which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*A*	document member of the same patent family
*O* document referring to an oral disclosure, use, exhibition or other means		
*P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

05 AUGUST 1994

Date of mailing of the international search report

19 AUG 1994

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JOHN D. ULM

Telephone No. (703) 308-0196

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US94/04651

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	The Journal of Biological Chemistry, Volume 264, issued 05 December 1989, B.D. Lemire et al., "The Mitochondrial Targeting Function of Randomly Generated Peptide Sequences Correlates with Predicted Helical Amphiphilicity", pages 20206-20215, see entire document.	1-27, 39-41, 45-63
Y	NATURE Volume 354, issued 07 November 1991, R.A. Houghten et al., "Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery", pages 84-86, see entire document.	1-27, 39-41, 45-63
Y	Nucleic Acids Research, Volume 16, Number 11, issued 10 June 1988, K.T. Min et al., "Search for the optimal sequence of the ribosomal binding site by random oligonucleotide-directed mutagenesis", pages 5075-5099, see entire document.	28-38, 42-44, 64-72
Y	Nucleic Acids Research, Volume 16, Number 15, issued 11 August 1988, A.R. Oliphant et al., "Defining the consensus sequences of E. coli promoter elements by random selection", pages 7673-7683, see entire document.	28-38, 42-44, 64-72